# PCT

## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

| (51) International Patent Classification [7] : | | (11) International Publication Number: | **WO 00/18906** |
|---|---|---|---|
| C12N 15/10, 15/11, 5/10, 7/04, A61K 39/21, C07B 61/00, G06F 19/00, B01J 19/00 // C12N 15/12, 15/55, 15/49 | **A2** | (43) International Publication Date: | 6 April 2000 (06.04.00) |

| | |
|---|---|
| (21) International Application Number: PCT/US99/22588 | (81) **Designated States:** AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG). |
| (22) International Filing Date: 28 September 1999 (28.09.99) | |
| (30) **Priority Data:** | |
| 60/102,362    29 September 1998 (29.09.98)   US | |
| 60/117,729    29 January 1999 (29.01.99)   US | |
| 60/118,813    5 February 1999 (05.02.99)   US | |
| 60/141,049    24 June 1999 (24.06.99)   US | |
| | **Published** |
| (71) **Applicant** *(for all designated States except US)*: MAXYGEN, INC. [US/US]; 515 Galveston Drive, Redwood City, CA 94063 (US). | *Without international search report and to be republished upon receipt of that report.* |
| (72) **Inventors; and** | |
| (75) **Inventors/Applicants** *(for US only)*: PATTEN, Phillip, A. [US/US]; Apartment 506, 2680 Fayette Drive, Mountain View, CA 94040 (US). LIU, Lu [US/US]; 519 Skiff Circle, Redwood City, CA 94056 (US). STEMMER, Willem, P., C. [NL/US]; 108 Kathy Court, Los Gatos, CA 95030 (US). | |
| (74) **Agents:** QUINE, Jonathan, Alan; Law Offices of Jonathan Alan Quine, P.O. Box 458, Alameda, CA 94501 (US) et al. | |

(54) Title: SHUFFLING OF CODON ALTERED GENES

(57) **Abstract**

Methods of recombining codon–altered libraries of nucleic acids are provided. The nucleic acids can include conservative or non–conservative modifications of coding sequences, in addition to codon alterations, as compared with wild–type sequences. In addition to making new proteins, methods of generation vectors with reduced rates of reversion to wild–type and attenuated viruses are also provided.

## CROSS REFERENCE TO RELATED APPLICATIONS

This application is a non-provisional filing of USSN 60/102,362,

5       "SHUFFLING OF CODON ALTERED GENES," Attorney Docket No. 02-028500, by

Patten and Stemmer, filed 09/29/98, and 60/117,729, "SHUFFLING OF CODON ALTERED

GENES," Attorney Docket No. 02-028510, by Patten and Stemmer, filed January 29, 1999.

The application is also related to USSN 60/118,813 "OLIGONUCLEOTIDE MEDIATED

NUCLEIC ACID RECOMBINATION," by Crameri et al., Attorney Docket Number 02-296,

10      filed February 5, 1999; and USSN 60/141,049 "OLIGONUCLEOTIDE MEDIATED

NUCLEIC ACID RECOMBINATION," by Crameri et al., Attorney Docket Number 02-296-

1, filed June 24, 1999.

## BACKGROUND

The genetic code is highly degenerate. Every DNA/RNA triplet (codon)

15      encoding an amino acid can typically be altered, with the exception of ATG/AUG (coding for

methionine) and TGG/UGG (coding for Tryptophan), without altering the sequence of the

protein encoded by the corresponding nucleic acid sequence. Roughly, on average (the

distribution of amino acids varies from protein to protein), each coding triplet can be

substituted about 3 different ways, since there are 61 codons encoding 20 amino acids (there

20      are 3 additional triplets encoding stop codons, for a total of 64 codons encoding 20 amino

acids). This represents a possible sequence diversity of approximately $3^n$ possible sequences

which encode a given protein, where n is the length of the protein in amino acids. As can

easily be seen, for proteins of even modest length, the number of possible nucleic acids which

can encode the protein exceeds the number of physical particles in the universe (estimated at

25      about $10^{80}$ particles).

This tremendous potential coding sequence space for individual proteins has

interesting evolutionary implications. For example, hypermutable viruses such as HIVs and

other retroviruses typically stay one step ahead of the host immune system by accumulating

non-random mutations based, in part, upon the particular codons used to encode recognition

30      molecules, e.g., in the envelope portion of the virus. The mutations are non-random because

viruses are selected for the ability to mutate to forms which are not quickly recognized by the

host immune system. A consequence of this is that viruses are selected to have a non-random set of codons encoding, e.g., envelope proteins, allowing the viruses to shift forms rapidly by making, e.g., specific point mutations to generate specific alterations in protein structure.

Codon use is also non-random within species. By preferentially making a subset of all possible t-RNAs, cells may conserve energy, and can optimize, or even regulate, the efficiency of cellular translation systems. This fact has long been recognized empirically, often allowing investigators initially to determine the reading frame of a given nucleic acid sequence simply by consideration of the codons resulting from different potential reading frames. One consequence of this "species codon bias" is that proteins within a species have a limited set of possible mutations that can arise as a consequence of, e.g., point mutation. This limits the possible evolution rate of proteins.

In addition to the diversity of nucleic acid coding sequences which encode any given protein, it is now clear that protein sequences are, themselves, quite degenerate. Often, many of the amino acid residues constituting a protein may be substituted for structurally similar amino acid units without significantly changing the tertiary structure of the protein. Thus, it may be difficult to determine which residues to modify or to improve desirable properties of a protein.

For proteins which are commercially valuable, it would be desirable to be able to gain access to a mutational spectrum which is different than that of the native protein. The present invention provides this, and many other features, that will be apparent upon complete review of the following.

## SUMMARY OF THE INVENTION

The present invention provides methods of accessing a completely different mutational spectrum for a selected protein than is available in the naturally occurring nucleic acid encoding the protein. This increases the type and rate of forced evolution for the selected protein, allowing for rapid improvement of any detectable characteristic of the protein. In the methods, nucleic acids are synthesized with altered codon usage, and/or which encode one or several amino acid residue changes as compared to the selected protein, where the amino acid and codon usage changes can be conservative or non-conservative. The resulting codon/amino acid modified nucleic acid(s) are recombined using DNA shuffling techniques with either the native nucleic acid, or with each other (or both), typically

2

using recursive shuffling methods. The nucleic acids or the encoded protein are then screened for a desirable property.

Thus, the invention provides methods of making codon altered nucleic acids. In the methods, a first nucleic acid sequence encoding a first polypeptide sequence is

5    selected. A plurality of codon altered nucleic acid sequences, each of which encode the first polypeptide, or a modified form thereof, are then selected (e.g., a library of codon altered nucleic acids can be selected in a biological assay which recognizes library components or activities), and the plurality of codon altered nucleic acid sequences is recombined to produce a target codon altered nucleic acid encoding a second protein. The target codon altered

10   nucleic acid is then screened for a detectable functional or structural property, optionally including comparison to the properties of the first polypeptide. The goal of such screening is to identify a polypeptide that has a structural or functional property equivalent or superior to the first polypeptide. A nucleic acid encoding such a polypeptide can be used in essentially any procedure desired, including introducing the target codon altered nucleic acid into a cell,

15   vector, virus, attenuated virus (e.g., as a component of a vaccine or immunogenic composition), transgenic organism, or the like.

Kits and compositions for practicing the methods are also provided, including one or more of: cell recombination mixtures and substrates (e.g., nucleic acids with altered codon usage), containers, instructional material for practicing the methods, or the like.

20                          **BRIEF DESCRIPTION OF THE FIGURES**
Figure 1 is a nucleic acid/amino acid sequence of a part of the monkey EPO gene, which is similar to the human EPO gene.

Figure 2 shows an example of a codon altered EPO nucleic acid sequence.

Figure 3 shows an alignment of naturally occurring EPOs.

25   Figure 4 is a schematic of the human EPO wobble sequence space.

Figure 5 is a schematic of Mammalian EPO Family-Wobble Sequence Space.

Figure 6 is a sequence alignment of G-CSF homologs, with species information.

Figure 7 is a sequence alignment of G-CSF homologs, with differences broken

30   out.

3

Figure 8 is a sequence alignment showing the hydrophobic core residues of human G-CSF (blacked out).

Figure 9 is a schematic showing the shuffling strategy for G-CSF.

Figure 10 is a list of oligos used to make a codon altered alkaline phosphatase.

Figure 11 is a map of oligos used to make a codon altered alkaline phosphatase.

Figure 12 is a schematic of vaccination with evolution defective viruses.

Figure 13 is a schematic of different mutations that result from different codon types for ser, arg, and leu.

Figure 14 is a schematic of vaccination with evolution defective viruses.

Figure 15 is a schematic of vaccination with evolution defective viruses showing sophisticated versus non-sophisticated "mutant clouds."

Figure 16, panels A-C show results of single mutations of different codons for ser, arg, and leu.

Fig. 17 is a schematic of protein evolution with expanded mutation spectra.

Fig. 18, panels A-D show codon altered forms of Env.

Fig. 19 is a list of oligos in one application for synthesis of HIV Env.

## DEFINITIONS

Unless clearly indicated to the contrary, the following definitions supplement definitions of terms known in the art.

As used herein, a "recombinant" nucleic acid is a nucleic acid produced by recombination between two or more nucleic acids, or any nucleic acid made by an *in vitro* or artificial process. The term "recombinant" when used with reference to a cell indicates that the cell comprises (and optionally replicates) a heterologous nucleic acid, or expresses a peptide or protein encoded by a heterologous nucleic acid. Recombinant cells can contain genes that are not found within the native (non-recombinant) wild-type form of the cell. Recombinant cells can also contain genes found in the native form of the cell where the genes are modified and re-introduced into the cell by artificial means. The term also encompasses cells that contain a nucleic acid endogenous to the cell that has been artificially modified without removing the nucleic acid from the cell; such modifications include those obtained by gene replacement, site-specific mutation, chimeraplasty, and related techniques.

4

A "codon altered" nucleic acid is a first nucleic acid that encodes a first polypeptide similar or identical to a naturally occurring polypeptide encoded by a naturally occurring nucleic acid, where the first nucleic acid utilizes a plurality of codons to encode the first polypeptide, which differ from the codons of the naturally occurring nucleic acid that

5      encode the naturally occurring polypeptide.

A "nucleic acid sequence" refers to either a nucleic acid (e.g., RNA, DNA or modified form thereof, in isolated, recombinant or native form) or to a representation of the nucleic acid such as a sequence of letters indicating the primary structure (sequence) of the nucleic acid.

10     A "polypeptide sequence" refers to either a polypeptide (or modified form thereof, in isolated, recombinant or native form) or to a representation of the polypeptide such as a sequence of letters or other character string information indicating the primary structure (amino acid sequence) of the polypeptide.

A "modified form" of a reference polypeptide is a target polypeptide which

15     has a similar, but not identical, sequence to the reference polypeptide. The sequence of the target polypeptide can differ from the reference polypeptide by conservative or non-conservative substitutions of the reference polypeptide sequence. As noted in more detail, *supra*, different nucleic acids encoding different target polypeptides having different non-conservative substitutions relative to the reference polypeptide can be recombined to produce

20     a recombined nucleic acid encoding a target polypeptide more similar to the reference polypeptide.

A "plurality of forms" of a selected nucleic acid refers to a plurality of homologs of the nucleic acid. The homologs can be from naturally occurring homologs (e.g., two or more homologous genes, or derivatives thereof) or by artificial synthesis of one or

25     more nucleic acids having related sequences, or by modification of one or more nucleic acid to produce related nucleic acids. Nucleic acids are homologous when they are derived, naturally or artificially, from a common ancestor sequence. During natural evolution, this occurs when two or more descendent sequences diverge from a parent sequence over time, i.e., due to mutation and natural selection. Under artificial conditions, divergence occurs,

30     e.g., in one of two ways. First, a given sequence can be artificially recombined with another sequence, as occurs, e.g., during typical cloning, to produce a descendent nucleic acid.

Alternatively, a nucleic acid can be synthesized *de novo*, by synthesizing a nucleic acid which varies in sequence from a given parental nucleic acid sequence.

When there is no explicit knowledge about the ancestry of two nucleic acids, homology is typically inferred by sequence comparison between two sequences. Where two

5    nucleic acid sequences show sequence similarity it is inferred that the two nucleic acids share a common ancestor. The precise level of sequence similarity required to establish homology varies in the art depending on a variety of factors. For purposes of this disclosure, two sequences are considered homologous where they share sufficient sequence identity to allow recombination to occur between two nucleic acid molecules, or when codon changes can be

10   made which would result in two or more nucleic acids having the ability to recombine. Typically, nucleic acids require regions of close similarity spaced roughly the same distance apart to permit recombination to occur.

The terms "identical" or percent "identity," in the context of two or more nucleic acid or polypeptide sequences, refer to two or more sequences or subsequences that

15   are the same or have a specified percentage of amino acid residues or nucleotides that are the same, when compared and aligned for maximum correspondence, as measured using one of the sequence comparison algorithms described below (or other algorithms available to persons of skill) or by visual inspection.

The phrase "substantially identical," in the context of two nucleic acids or

20   polypeptides refers to two or more sequences or subsequences that have at least about 40%, 50%, 60%, or preferably about 70% or 80% or more, or most preferably 90-95% nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using one of the following sequence comparison algorithms or by visual inspection. Such "substantially identical" sequences are typically considered to be homologous.

25   Preferably, the "substantial identity" exists over a region of the sequences that is at least about 50 residues in length, more preferably over a region of at least about 100 residues, and most preferably the sequences are substantially identical over at least about 150 residues, or over the full length of the two sequences to be compared.

For sequence comparison and homology determination, typically one

30   sequence acts as a reference sequence to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer,

6

subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

5            Optimal alignment of sequences for comparison can be conducted, e.g., by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and

10    TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by visual inspection (*see generally*, Ausubel et al., *infra*).

       One example of algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul *et al., J. Mol. Biol.* 215:403-410 (1990). Software for performing BLAST analyses is publicly

15    available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score

20    threshold (Altschul et al., *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for

25    mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm

30    parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation

(E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff (1989) *Proc. Natl. Acad. Sci. USA* 89:10915).

5          In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.,* Karlin & Altschul (1993) *Proc. Nat'l. Acad. Sci. USA* 90:5873-5787). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid

10       sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.1, more preferably less than about 0.01, and most preferably less than about 0.001.

          Another indication that two nucleic acid sequences are substantially identical/

15       homologous is that the two molecules hybridize to each other under stringent conditions. The phrase "hybridizing specifically to," refers to the binding, duplexing, or hybridizing of a molecule only to a particular nucleotide sequence under stringent conditions, including when that sequence is present in a complex mixture (*e.g.,* total cellular) DNA or RNA. "Bind(s) substantially" refers to complementary hybridization between a probe nucleic acid and a

20       target nucleic acid and embraces minor mismatches that can be accommodated by reducing the stringency of the hybridization media to achieve the desired detection of the target polynucleotide sequence.

          "Stringent hybridization conditions" and "stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and

25       northern hybridizations are sequence dependent, and are different under different environmental parameters. Longer sequences and sequences with higher G:C content remain hybridized at higher temperatures (or at lower salt). An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993) *Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes* part I chapter 2 "Overview of

30       principles of hybridization and the strategy of nucleic acid probe assays," Elsevier, New York.

Generally, highly stringent hybridization and wash conditions are selected to be about 5 °C lower than the thermal melting point ($T_m$) for the specific sequence at a defined ionic strength and pH. Typically, under "stringent conditions" a probe will hybridize to its target subsequence, but not to unrelated (non-homologous) sequences.

5          The $T_m$ is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the $T_m$ for a particular probe. An example of stringent hybridization conditions for hybridization of complementary nucleic acids which have more than 100 complementary residues on a filter in a Southern or northern blot is 50% formamide with 1

10        mg of heparin at 42 °C, with the hybridization being carried out overnight. An example of highly stringent wash conditions is 0.15M NaCl at 72 °C for about 15 minutes. An example of stringent wash conditions is a 0.2x SSC wash at 65 °C for 15 minutes (see, Sambrook, infra., for a description of SSC buffer). Often, a high stringency wash is preceded by a low stringency wash to remove background probe signal. An example medium stringency wash

15        for a duplex of, e.g., more than 100 nucleotides, is 1x SSC at 45 °C for 15 minutes. An example low stringency wash for a duplex of, e.g., more than 100 nucleotides, is 4-6x SSC at 40 °C for 15 minutes. For short probes (e.g., about 10 to 50 nucleotides), stringent conditions typically involve salt concentrations of less than about 1.0 M Na ion, typically about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3, and the temperature is typically

20        at least about 40 °C. Stringent conditions can also be achieved with the addition of destabilizing agents such as formamide. In general, a signal to noise ratio of 2x (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization. If the signal to noise ratio is less than 2x binding of an unrelated probe (e.g., a nucleic acid encoding a non-homologous protein), the nucleic acids at

25        issue do not hybridize under stringent conditions. Similarly, if the signal to noise ratio is less than 25% as high as that observed for a perfectly matched probe under stringent conditions, the nucleic acids do not "hybridize under stringent conditions" as that term is used herein. This does not apply to highly stringent conditions, as the stringency can theoretically be increased until only a perfectly matched probe will hybridize.

30        In one example hybridization procedure, a target nucleic acid to be probed is blotted onto a filter by any conventional method. An unrelated nucleic acid such as a plasmid

9

vector (assuming that the target nucleic acid has no homology with the target nucleic acid) is also blotted, in approximately equal amounts onto the filter. The filter is probed with a labeled probe complementary to the target nucleic acid. The experiment is repeated at gradually increasing stringency of hybridization and wash conditions until signal from the

5      hybridization of the labeled probe to the complementary target is 10-100X as high as to the unrelated plasmid vector nucleic acid. Once these conditions are determined as described above, a test nucleic acid is probed under the same conditions as the target. If signal from the labeled probe is 25% as high or higher than the signal from binding of the probe to the target, the test nucleic acid "hybridizes under stringent conditions" to the probe. If the signal is less

10     than 25% as high, the test nucleic acid does not hybridize under stringent conditions to the probe.

               Nucleic acids which do not hybridize to each other under stringent conditions are still recognizable as variant forms of a nucleic acid when the polypeptides they encode are substantially identical. This occurs, e.g., when a copy of a nucleic acid is created using

15     the maximum codon degeneracy permitted by the genetic code. Such nucleic acids are not functionally equivalent, as described in detail herein, due to differences in mRNA folding, alterations of regulatory sequences and the like.

               Another indication that two nucleic acid sequences or polypeptides are variant forms is that the polypeptide encoded by the first nucleic acid is immunologically cross

20     reactive with the polypeptide encoded by the second nucleic acid, as tested by polyclonal antisera generated to the first polypeptide. Thus, a polypeptide is typically substantially identical to a second polypeptide, for example, where the two peptides differ only by conservative substitutions.

               "Conservatively modified variations" of a particular polynucleotide sequence

25     are those polynucleotide variations that encode identical or essentially identical amino acid sequences, or where the polynucleotide does not encode an amino acid sequence, which encode essentially identical sequences. Because of the degeneracy of the genetic code, a large number of functionally identical nucleic acids encode any given polypeptide. For instance, the codons CGU, CGC, CGA, CGG, AGA, and AGG all encode the amino acid arginine.

30     Thus, at every position where an arginine is specified by a codon, the codon can be altered to any of the corresponding codons described without altering the encoded polypeptide. Such

nucleic acid variations are "silent variations," which are one species of "conservatively modified variations." Every polynucleotide sequence described herein which encodes a polypeptide also optionally describes every possible silent variation, except where otherwise noted. One of skill will recognize that each codon in a nucleic acid (except AUG, which is ordinarily the codon for methionine, and TGG, which is ordinarily the codon for tryptophan) can be modified to yield a peptide which is structurally identical.

Furthermore, one of skill will recognize that individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids (typically less than 5%, more typically less than 1%) in an encoded sequence are "conservatively modified variations" where the alterations result in the substitution of an amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art. The following five groups each contain amino acids that are conservative substitutions for one another:

Aliphatic: Glycine (G), Alanine (A), Valine (V), Leucine (L), Isoleucine (I); Aromatic: Phenylalanine (F), Tyrosine (Y), Tryptophan (W); Sulfur-containing: Methionine (M), Cysteine (C);Basic: Arginine (R), Lysine (K), Histidine (H); Acidic: Aspartic acid (D), Glutamic acid (E), Asparagine (N), Glutamine (Q). *See also*, Creighton (1984) *Proteins*, W.H. Freeman and Company. In addition, individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids in an encoded sequence are also "conservatively modified variations." Sequences that differ by conservative variations are generally homologous.

The term "isolated", when applied to a nucleic acid or protein, denotes that the nucleic acid or protein is essentially free of other cellular or other components (e.g., library components) with which it is associated in the natural state.

The term "nucleic acid" refers to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form. Unless specifically limited, the term encompasses nucleic acids containing known analogues of natural nucleotides which have similar binding properties as the reference nucleic acid and are metabolized in a manner similar to naturally occurring nucleotides. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (*e.g.* degenerate codon substitutions) and complementary sequences and as well as the sequence

11

explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzer *et al.* (1991) *Nucleic Acid Res. 19*: 5081; Ohtsuka *et al.* (1985) *J. Biol. Chem. 260*: 2605-2608; Cassol *et al.* (1992) ;

5    Rossolini *et al.* (1994) *Mol. Cell. Probes 8*: 91-98). The term nucleic acid is generic to the terms "gene", "DNA," "cDNA", "oligonucleotide," "RNA," "mRNA," and the like.

"Nucleic acid derived from a gene" refers to a nucleic acid for whose synthesis the gene, or a subsequence thereof, has ultimately served as a template. Thus, an mRNA, a cDNA reverse transcribed from an mRNA, an RNA transcribed from that cDNA, a

10   DNA amplified from the cDNA, an RNA transcribed from the amplified DNA, *etc.*, are all derived from the gene and detection of such derived products is indicative of the presence and/or abundance of the original gene and/or gene transcript in a sample.

A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is

15   operably linked to a coding sequence if it increases the transcription of the coding sequence.

A "recombinant expression cassette" or simply an "expression cassette" is a nucleic acid construct, generated recombinantly or synthetically, with nucleic acid elements that are capable of effecting expression of a structural gene in hosts compatible with such sequences. Expression cassettes include at least promoters and optionally, transcription

20   termination signals. Typically, the recombinant expression cassette includes a nucleic acid to be transcribed (*e.g.*, a nucleic acid encoding a desired polypeptide), and a promoter. Additional factors necessary or helpful in effecting expression may also be used as described herein. For example, an expression cassette can also include nucleotide sequences that encode a signal sequence that directs secretion of an expressed protein from the host cell.

25   Transcription termination signals, enhancers, and other nucleic acid sequences that influence gene expression, can also be included in an expression cassette.

## DETAILED DISCUSSION OF THE INVENTION
In the present invention, the sequence diversity of substrates for DNA shuffling procedures is increased by using codon-altered nucleic acids as templates and/or by

30   using templates that encode proteins with conservative or non-conservative amino acid modifications as compared to a selected wild-type protein.

These codon altered nucleic acids can be chemically synthesized (e.g., using standard artificial synthetic protocols, e.g., those typically used by commercial sources from which nucleic acids can be ordered), or can be made using any of a variety of methods herein of available to one of skill. For example, oligonucleotide fragments can be made which

5    correspond to a codon altered nucleic acid which is desired using standard synthetic methods, followed by polymerase and/or ligase mediated oligonucleotide ligation/recombination protocols to generate full-length nucleic acids.

The combination of codon usage modifications and coding modifications can be extensive enough to reduce or, under stringent conditions, even eliminate the hybridization

10   of the codon-altered nucleic acids to a nucleic acid which naturally encodes the selected protein. This dramatically alters the mutations which result from possible single nucleotide mutations, providing access to greater diversity for DNA shuffling protocols.

In addition, the recombination and selection of such nucleic acids during DNA shuffling procedures can result not only in access to a different set of possible mutations, but

15   can also result in modified forms of transcriptional or translational regulation, alterations in nucleic acid localization, mRNA stability and the like. Furthermore, the modified hybridization properties of codon altered nucleic acids leads to alterations in the ability of the nucleic acids to hybridize with potential recombination partners, altering, and ultimately increasing, the available recombination diversity during shuffling.

20   Furthermore, "family shuffling" using codon-altered substrates even further increases the possible sequence diversity of the starting materials for recombination. As currently practiced, family shuffling methods involve shuffling nucleic acids encoding sequence variants of a given protein (e.g., species or allele homologs). In the present methods, this procedure is modified by generating codon-altered versions of the sequence

25   variants to access additional molecular diversity during recombination. Additional diversity is achieved by conservatively and non-conservatively modifying the starting nucleic acids to encode non-naturally occurring sequence variants. Family shuffling can be performed even using homologs of relatively low identity. In such cases, codons may be changed in one or more of the family members to increase the level of identity between the members, thereby

30   increasing their ability to recombine using the methods of this invention.

Gene shuffling and family shuffling provide two of the most powerful methods available for improving and "migrating" (gradually changing the type of reaction, substrate or activity of a selected protein such as an enzyme, or regulation or structure of an expressed component) the functions of proteins. In family shuffling, homologous sequences,

5     e.g., from different species, chromosomal positions, or due to synthetic alteration, are recombined. In gene shuffling, a single sequence is mutated or otherwise altered and then recombined.

The generation and screening of high quality shuffled libraries provides for DNA shuffling (or "directed evolution"). The availability of appropriate high-throughput

10    analytical chemistry to screen the libraries permits integrated high-throughput shuffling and screening of the libraries to achieve a desired activity.

In one significant embodiment, oligonucleotides for constructing codon-modified nucleic acids are designed in a computer ("in silico"). Predicted codon-modified recombinant nucleic acids can also be determined in silico, i.e., essentially as taught in

15    Selifonov and Stemmer "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" filed 02/05/1999, USSN 60/118854.

Furthermore, rather than generating codon-modified nucleic acids as substrates for recombination, families of nucleic acids can be recombined simply by

20    appropriate selection of the relevant oligonucleotides which are used in gene reconstruction methods to produce recombinant nucleic acids, i.e., by using codon-modified nucleic acid oligonucleotides as discussed herein in conjunction with family oligonucleotide-mediated shuffling methods, e.g., as taught in Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed February 5, 1999, USSN 60/118,813 and

25    Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed June 24, 1999, USSN 60/141,049. The technique can be used to recombine homologous or even non-homologous nucleic acid sequences; in the context of the present invention, oligonucleotides corresponding to families of codon-modified nucleic acids are shuffled.

The present invention provides significant advantages over previously used

30    methods for optimization of genes. For example, DNA shuffling of codon modified nucleic acids can result in optimization of a desirable property even in the absence of a detailed

understanding of the mechanism by which the particular property is mediated. In addition, entirely new properties can be obtained upon shuffling of codon modified DNAs, i.e., shuffled DNAs can encode polypeptides or RNAs with properties entirely absent in the parental DNAs which are shuffled. Thus, by modifying the codon usage and/or encoded

5     amino acids of the relevant gene or other nucleic acid, molecular diversity is accessed and sequences can be shuffled to obtain desired, including entirely new, properties.

In general, sequence recombination can be achieved in many different formats and permutations of formats, as described in further detail below.

The targets for modification vary in different applications, as does the

10    property sought to be acquired or improved. Examples of candidate targets for acquisition of a property or improvement in a property include genes that encode proteins which have enzymatic or therapeutic or other commercially useful activities. A more extensive listing is found *supra*; however, even this list is not intended to be limiting, as essentially any nucleic acid can be codon modified and shuffled, using one or more of the processes herein.

15    Shuffling methods use at least two variant forms of a starting target (the variant forms can be nucleic acids, or representations thereof, e.g., as character strings in a computer program). The variant forms of candidate codon-altered substrates can show substantial sequence or secondary structural similarity with each other, but they should also differ in at least one and preferably at least two positions. The initial diversity between forms

20    can be the result of natural variation, e.g., the different variant forms (homologs) are obtained from different individuals or strains of an organism, or constitute related sequences from the same organism (e.g., allelic variations), or constitute homologs from different organisms (interspecific variants), or constitute artificial homologs, e.g., codon-altered nucleic acids encoding the same or a similar protein. Any or all of these sequences can represent or

25    include codon altered nucleic acids.

Initial diversity can also be induced, e.g., the variant forms can be generated by error-prone transcription, such as an error-prone PCR or use of a polymerase which lacks proof-reading activity (*see,* Liao (1990) *Gene* 88:107-111), of the first variant form, or, by replication of the first form in a mutator strain (mutator host cells are discussed in further

30    detail below, and are generally well known). The initial diversity between substrates is greatly augmented in subsequent steps of recombination for library generation.

15

A mutator strain can include any mutants in any organism impaired in the functions of mismatch repair. These include mutant gene products of mutS, mutT, mutH, mutL, ovrD, dcm, vsr, umuC, umuD, sbcB, recJ, etc. The impairment is achieved by genetic mutation, allelic replacement, selective inhibition by an added reagent such as a small

5      compound or an expressed antisense RNA, or other techniques. Impairment can be of the genes noted, or of homologous genes in any organism. The properties or characteristics that can be acquired or improved vary widely, and, of course depend on the choice of substrate.

At least two variant forms of a nucleic acid, e.g., which can confer a desired activity or which can be recombined to produce a desired activity, are recombined to produce

10    a library of recombinant nucleic acids. The library is then screened to identify at least one recombinant nucleic acid that is optimized for the particular property or properties of interest.

Often, improvements are achieved after one round of recombination and selection. However, recursive sequence recombination can be employed to achieve still further improvements in a desired property, or to bring about new (or "distinct") properties.

15    Recursive sequence recombination entails successive cycles of recombination to generate molecular diversity. That is, one creates a family of nucleic acid molecules showing some sequence identity to each other but differing due to the presence of mutations. In any given cycle, recombination can occur *in vivo* or *in vitro*, intracellularly or extracellularly. Furthermore, diversity resulting from recombination can be augmented in any cycle by

20    applying known methods of mutagenesis (*e.g.*, error-prone PCR or cassette mutagenesis) to either the substrates or products for recombination. In general, however, a single cycle of DNA shuffling of codon-altered nucleic acids provides for generation of surprisingly effective nucleic acids. Accordingly, while recursive approaches to shuffling can be used, single cycle recombination is also preferred. Typically, 2, 3, 4, 5, or even 10 or more cycles

25    of recombination can be performed, each cycle optionally comprising one or more selection steps.

A recombination cycle is usually followed by at least one cycle of screening or selection for molecules having a desired property or characteristic. If a recombination cycle is performed *in vitro*, the products of recombination, *i.e.*, recombinant segments, are

30    sometimes introduced into cells before the screening step. Recombinant segments can also be linked to an appropriate vector or other regulatory sequences before screening.

Alternatively, products of recombination generated *in vitro* are sometimes packaged in viruses (e.g., bacteriophage) before screening. If recombination is performed *in vivo*, recombination products can sometimes be screened in the cells in which recombination occurred. In other applications, recombinant segments are extracted from the cells, and

5    optionally packaged as viruses, before screening.

The nature of screening or selection depends on what property or characteristic is to be acquired or the property or characteristic for which improvement is sought, and many examples are discussed below. It is not usually necessary to understand the molecular basis by which particular products of recombination (recombinant segments) have

10   acquired new or improved properties or characteristics relative to the starting substrates. For example, a gene can have many component sequences, each having a different intended role (*e.g.*, coding sequences, regulatory sequences, targeting sequences, stability-conferring sequences, subunit sequences and sequences affecting integration). Each of these component sequences can be varied and recombined simultaneously. Screening/selection can then be

15   performed, for example, for recombinant segments that have increased ability to confer activity upon a cell without the need to attribute such improvement to any of the individual component sequences of the vector.

Depending on the particular screening protocol used for a desired property, initial round(s) of screening can sometimes be performed using bacterial cells due to high

20   transfection efficiencies and ease of culture. However, bacterial expression is often not practical or desired, and yeast, fungal or other eukaryotic systems are also used for library expression and screening. Similarly, other types of screening which are not amenable to screening in bacterial or simple eukaryotic library cells, are performed in cells selected for use in an environment close to that of their intended use. Final rounds of screening can be

25   performed in the precise cell type of intended use.

If further improvement in a property is desired, at least one and usually a collection of recombinant segments surviving a first round of screening/selection are subject to a further round of recombination. These recombinant segments can be recombined with each other or with exogenous segments representing the original substrates or further variants

30   thereof. Again, recombination can proceed *in vitro* or *in vivo*. If the previous screening step identifies desired recombinant segments as components of cells, the components can be

17

subjected to further recombination *in vivo*, or can be subjected to further recombination *in vitro*, or can be isolated before performing a round of *in vitro* recombination. Conversely, if the previous screening step identifies desired recombinant segments in naked form or as components of viruses, these segments can be introduced into cells to perform a round of *in vivo* recombination. The second round of recombination, irrespective how performed, generates further recombinant segments which encompass additional diversity that is present in recombinant segments resulting from a previous round (or from multiple previous rounds, e.g., where the process is iteratively repeated).

The second round of recombination can be followed by a further round of screening/selection according to the principles discussed above for the first round. The stringency of screening/selection can be increased between rounds. Also, the nature of the screen and the property being screened for can vary between rounds if improvement in more than one property is desired or if acquiring more than one new property is desired. Additional rounds of recombination and screening can then be performed until the recombinant segments have sufficiently evolved to acquire the desired new or improved property or function.

The practice of this invention involves the construction of recombinant nucleic acids and the expression of genes in transfected host cells. Molecular cloning techniques to achieve these ends are known in the art. A wide variety of cloning and *in vitro* amplification methods suitable for the construction of recombinant nucleic acids such as expression vectors are well-known to persons of skill. General texts which describe molecular biological techniques useful herein, including mutagenesis, include Berger and Kimmel, Guide to Molecular Cloning Techniques, Methods in Enzymology volume 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook et al., Molecular Cloning - A Laboratory Manual (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989 ("Sambrook") and Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 1998) ("Ausubel")). Methods of transducing cells, including plant and animal cells, with nucleic acids are generally available, as are methods of expressing proteins encoded by such nucleic acids. In addition to Berger, Ausubel and Sambrook, useful general references for culture of animal cells include Freshney (Culture of

Animal Cells, a Manual of Basic Technique, third edition Wiley- Liss, New York (1994)) and
the references cited therein, Humason (Animal Tissue Techniques, fourth edition W.H.
Freeman and Company (1979)) and Ricciardelli, et al., In Vitro Cell Dev. Biol. 25:1016-1024
(1989). References for plant cell cloning, culture and regeneration include Payne et al.

5      (1992) Plant Cell and Tissue Culture in Liquid Systems John Wiley & Sons, Inc. New York,
NY (Payne); and Gamborg and Phillips (eds) (1995) Plant Cell, Tissue and Organ Culture;
Fundamental Methods Springer Lab Manual, Springer-Verlag (Berlin Heidelberg New York)
(Gamborg). A variety of Cell culture media are described in Atlas and Parks (eds) The
Handbook of Microbiological Media (1993) CRC Press, Boca Raton, FL (Atlas). Additional

10     information for plant cell culture is found in available commercial literature such as the Life
Science Research Cell Culture Catalogue (1998) from Sigma- Aldrich, Inc (St Louis, MO)
(Sigma-LSRCCC) and, e.g., the Plant Culture Catalogue and supplement (1997) also from
Sigma-Aldrich, Inc (St Louis, MO) (Sigma-PCCS).

       Examples of techniques sufficient to direct persons of skill through *in vitro*
15     amplification methods, including the polymerase chain reaction (PCR), the ligase chain
reaction (LCR), Qβ-replicase amplification and other RNA polymerase mediated techniques
(*e.g.*, NASBA) are found in Berger, Sambrook, and Ausubel, *id.*, as well as in Mullis *et al.*,
(1987) U.S. Patent No. 4,683,202; PCR Protocols A Guide to Methods and Applications
(Innis *et al.* eds) Academic Press Inc. San Diego, CA (1990) (Innis); Arnheim & Levinson

20     (October 1, 1990) C&EN  36-47; The Journal Of NIH Research (1991) 3, 81-94; Kwoh *et al.*
(1989) Proc. Natl. Acad. Sci. USA 86, 1173; Guatelli *et al.* (1990) Proc. Natl. Acad. Sci.
USA 87, 1874; Lomell *et al.* (1989) J. Clin. Chem 35, 1826; Landegren *et al.*, (1988) Science
241, 1077-1080; Van Brunt (1990) Biotechnology 8, 291-294; Wu and Wallace, (1989) Gene
4, 560; Barringer *et al.* (1990) Gene 89, 117, and Sooknanan and Malek (1995)

25     Biotechnology 13: 563-564. Improved methods of cloning *in vitro* amplified nucleic acids
are described in Wallace *et al.*, U.S. Pat. No. 5,426,039. Improved methods of amplifying
large nucleic acids by PCR are summarized in Cheng *et al.* (1994) Nature 369: 684-685 and
the references therein, in which PCR amplicons of up to 40kb are generated. One of skill will
appreciate that essentially any RNA can be converted into a double stranded DNA suitable

30     for restriction digestion, PCR expansion and sequencing using reverse transcriptase and a
polymerase. *See*, Ausbel, Sambrook and Berger, *all supra.*

Oligonucleotides *e.g.*, for use in *in vitro* amplification/ gene reconstruction methods, for use as gene probes, or as shuffling targets (e.g., synthetic genes or gene segments) are typically synthesized chemically according to the solid phase phosphoramidite triester method, e.g., as described by Beaucage and Caruthers (1981), *Tetrahedron Letts.*,

5   22(20):1859-1862, *e.g.*, using an automated synthesizer, e.g., as described in Needham-VanDevanter *et al.* (1984) *Nucleic Acids Res.*, 12:6159-6168 or as is now practiced routinely in the art. Oligonucleotides can also be custom made and ordered from a variety of commercial sources known to persons of skill. Purification of oligonucleotides (e.g., using gel-purification methods) to improve the quality of synthesized oligonucleotides can be

10  particularly desirable in the processes herein to improve the quality of nucleic acid synthesis protocols.

As noted, essentially any nucleic acid can be custom ordered from any of a variety of commercial sources, such as The Midland Certified Reagent Company (merc@oligos.com), The Great American Gene Company (http://www.genco.com),

15  ExpressGen Inc. (www.expressgen.com), Operon Technoloigies Inc. (Alameda, CA) and many others. Similarly, peptides and antibodies can be custom ordered from any of a variety of sources, such as PeptidoGenic (pkim@ccnet.com), HTI Bio-products, inc. (http://www.htibio.com), BMA Biomedicals Ltd (U.K.), BioSynthesis, Inc., and many others.

20  <u>CODON AND AMINO ACID ALTERED LIBRARIES</u>

In the methods of the invention, libraries of codon altered nucleic acids can be made and recombined. The codon altered nucleic acids can also include differences in encoded amino acid sequences, which can be either conservative or non-conservative in nature. The codon altered nucleic acids can be derived from a single parental amino acid

25  sequence, or can be derived from a family of original sequences, *e.g.*, natural or synthetic homologous variants of a given sequence. Libraries can exist, *e.g.*, in pools or aliquots of cells, viral plaques, enzymatically synthesized pools or aliquots of nucleic acids, or chemically synthesisized pools of nucleic acids. Methods of making libraries of nucleic acids are available and taught, *e.g.*, in Berger, Sambrook and Ausubel, *supra*. In one embodiment,

30  a library as used in the invention comprises at least 2 nucleic acid sequences. In additional

20

embodiments, the libraries of this invention comprise at least 2, 5, 10, 100, 1000, or more nucleic acid sequences.

As applied to the invention, libraries are typically constructed with a high percentage of codons altered relative to an initial (e.g., wild type) nucleic acid. Codon usage

5      divergence for each of the codon altered nucleic acids can be 50%, 75%, or even 90% or more as compared to the first nucleic acid. This eliminates hybridization to the parental nucleic acid (and thereby inhibits recombination with the parental nucleic acid, a desirable feature in certain embodiments discussed below).

In several embodiments of this invention, codons are modified in members of

10     a gene family so as to increase the degree of identity between the members. In one such embodiment, the genes are homologous genes from different species. In such cases, the degree of nucleic acid identity may be lower than the degree of amino acid identity, at least in part, because of differences in codon usage between the species. In additional embodiments, the homologous genes represent different members of a gene family within a single species.

15     Such genes may encode functionally distinct members of a gene family that nevertheless share significant structural or functional similarity. In preferred embodiments, homologous genes are reverse translated into nucleic acid sequences, and the nucleic acid sequences are modified so as to increase the level of identity between them. Nucleic acids with the modified sequences can then be synthesized *in vitro*. In particularly preferred embodiments,

20     the modified nucleic acid sequences are at least as identical to each other as the original amino acid sequences.

Additional sequence diversity is provided by generating nucleic acids with non-overlapping non-conservative substitutions in each of the codon altered nucleic acids as compared to the first nucleic acid. This provides for reversion to wild-type upon

25     recombination, while optionally allowing for the incorporation of non-conservative changes to the sequence in the event that they produce a detectable improvement during screening.

Modification of the codons of one or more of the codon altered nucleic acids to provide one or more different hydrophobic core residue for an encoded polypeptide as compared to the first polypeptide is also provided. This modification of core amino acids

30     provides minor differences in encoded proteins, while changing the mutational spectrum of the resulting nucleic acid, thereby increasing sequence diversity.

In addition, due to the constraints of the translational machinery for a given cell, codon usage may need to be altered when expressed sequences are shuttled between different organisms (e.g., animal cells, plant cells, bacterial cells, etc.) for optimal expression. This produces a nucleic acid which encodes the same protein, but which, after typical forms

5     of point mutation, will access a different mutational diversity than the original form of the protein.

In one embodiment, phage libraries are made and recombined in mutator strains such as cells with mutant or impaired gene products of mutS, mutT, mutH, mutL, ovrD, dcm, vsr, umuC, umuD, sbcB, recJ, etc. The impairment is achieved by genetic

10    mutation, allelic replacement, selective inhibition by an added reagent such as a small compound or an expressed antisense RNA, or other techniques. High multiplicity of infection (MOI) libraries are used to infect the cells to increase recombination frequency. Additional strategies for making phage libraries and or for recombining DNA from donor and recipient cells are set forth in U.S. Pat. No. 5,521,077. Additional recombination strategies

15    for recombining plasmids in yeast are set forth in WO 97 07205.

The library to be made can be an *in vitro* set of molecules, or present in cells, phage or the like. Virtual libraries of nucleic acids generated in silico are also a feature of the invention (*see also*, Selifonov and Stemmer, *supra*). Generally, the library is screened to identify at least one recombinant nucleic acid that exhibits distinct or improved activity

20    compared to the parental nucleic acid or nucleic acids which are recombined. Additional details on making appropriate libraries are found below, *e.g.*, in the section entitled "Formats for Sequence Recombination."

TARGETS FOR CODON MODIFICATION AND SHUFFLING

Essentially any nucleic acid can be codon altered and shuffled. No attempt is

25    made herein to identify the hundreds of thousands of known nucleic acids. Common sequence repositories for known proteins include GenBank EMBL, DDBJ and the NCBI. Other repositories can easily be identified by searching the internet.

One class of preferred targets for activation includes nucleic acids encoding therapeutic proteins such as erythropoietin (EPO), insulin, peptide hormones such as human

30    growth hormone; growth factors and cytokines such as epithelial Neutrophil Activating Peptide-78, GROα/MGSA, GROβ, GROγ, MIP-1α, MIP-1β, MCP-1, epidermal growth

factor, fibroblast growth factor, hepatocyte growth factor, insulin-like growth factor, the interferons, the interleukins, keratinocyte growth factor, leukemia inhibitory factor, oncostatin M, PD-ECSF, PDGF, pleiotropin, SCF, c-kit ligand, VEGEF, G-CSF etc. Many of these proteins are commercially available (*See, e.g.,* the Sigma BioSciences 1997

5   catalogue and price list), and the corresponding genes are well-known.

Another class of preferred targets are transcriptional and expression activators. Example transcriptional and expression activators include genes and proteins that modulate cell growth, differentiation, regulation, or the like. Expression and transcriptional activators are found in prokaryotes, viruses, and eukaryotes, including fungi, plants, and animals,

10  including mammals, providing a wide range of therapeutic targets. It will be appreciated that expression and transcriptional activators regulate transcription by many mechanisms, e.g., by binding to receptors, stimulating a signal transduction cascade, regulating expression of transcription factors, binding to promoters and enhancers, binding to proteins that bind to promoters and enhancers, unwinding DNA, splicing pre-mRNA, polyadenylating RNA, and

15  degrading RNA. Expression activators include cytokines, inflammatory molecules, growth factors, their receptors, and oncogene products, e.g., interleukins (e.g., IL-1, IL-2, IL-8, etc.), interferons, FGF, IGF-I, IGF-II, FGF, PDGF, TNF, TGF-α, TGF-β, EGF, KGF, SCF/c-Kit, CD40L/CD40, VLA-4/VCAM-1, ICAM-1/LFA-1, and hyalurin/CD44; signal transduction molecules and corresponding oncogene products, e.g., Mos, Ras, Raf, and Met; and

20  transcriptional activators and suppressors, e.g., p53, Tat, Fos, Myc, Jun, Myb, Rel, and steroid hormone receptors such as those for estrogen, progesterone, testosterone, aldosterone, the LDL receptor ligand and corticosterone.

Similarly, proteins from infectious organisms for possible vaccine applications, described in more detail below, including infectious fungi, e.g., *Aspergillus*,

25  *Candida* species; bacteria, particularly *E. coli*, which serves a model for pathogenic bacteria, as well as medically important bacteria such as *Staphylococci* (e.g., *aureus*), *Streptococci* (e.g., *pneumoniae*), *Clostridia* (e.g., *perfringens*), *Neisseria* (e.g., *gonorrhoea*), *Enterobacteriaceae* (e.g., *coli*), *Helicobacter* (e.g., *pylori*), *Vibrio* (e.g., *cholerae*), *Capylobacter* (e.g., *jejuni*), *Pseudomonas* (e.g., *aeruginosa*), *Haemophilus* (e.g., *influenzae*),

30  *Bordetella* (e.g., *pertussis*), *Mycoplasma* (e.g., *pneumoniae*), *Ureaplasma* (e.g., *urealyticum*), *Legionella* (e.g., *pneumophila*), *Spirochetes* (e.g., *Treponema*, *Leptospira*, and *Borrelia*),

*Mycobacteria* (e.g., *tuberculosis, smegmatis*), *Actinomyces* (e.g., *israelii*), *Nocardia* (e.g.,

*asteroides*), *Chlamydia* (e.g., *trachomatis*), *Rickettsia, Coxiella, Ehrilichia, Rochalimaea,*

*Brucella, Yersinia, Fracisella,* and *Pasteurella*; protozoa such as sporozoa (e.g., *Plasmodia*),

*rhizopods* (e.g., *Entamoeba*) and flagellates (*Trypanosoma, Leishmania, Trichomonas,*

5     *Giardia,* etc.); viruses such as ( + ) RNA viruses (examples include Poxviruses e.g., *vaccinia*;

Picornaviruses, e.g. *polio*; Togaviruses, e.g., *rubella*; Flaviviruses, e.g., HCV; and

Coronaviruses), ( - ) RNA viruses (examples include Rhabdoviruses, e.g., VSV;

Paramyxoviruses, e.g., RSV; Orthomyxoviruses, e.g., influenza; Bunyaviruses; and

Arenaviruses), dsDNA viruses (Reoviruses, for example), RNA to DNA viruses, i.e.,

10    Retroviruses, e.g., especially HIV and HTLV, and certain DNA to RNA viruses such as

Hepatitis B virus.

           Other proteins relevant to non-medical uses, such as inhibitors of transcription

or toxins of crop pests e.g., insects, fungi, weed plants, and the like, are also preferred targets

for shuffling. Industrially important enzymes such as monooxygenases, proteases, nucleases,

15    and lipases are also preferred targets. As an example, subtilisin can be evolved by shuffling

codon altered forms of the gene for subtilisin (von der Osten et al., J. Biotechnol. 28:55-68

(1993) provide a subtilisin coding nucleic acid). Proteins which aid in folding such as the

chaperonins are also preferred.

           Preferred known genes suitable for codon alteration and shuffling also include

20    the following: Alpha-1 antitrypsin, Angiostatin, Antihemolytic factor, Apolipoprotein,

Apoprotein, Atrial natriuretic factor, Atrial natriuretic polypeptide, Atrial peptides, C-X-C

chemokines (e.g., T39765, NAP-2, ENA-78, Gro-a, Gro-b, Gro-c, IP-10, GCP-2, NAP-4,

SDF-1, PF4, MIG), Calcitonin, CC chemokines (e.g., Monocyte chemoattractant protein-1,

Monocyte chemoattractant protein-2, Monocyte chemoattractant protein-3, Monocyte

25    inflammatory protein-1 alpha, Monocyte inflammatory protein-1 beta, RANTES, I309,

R83915, R91733, HCC1, T58847, D31065, T64262), CD40 ligand, Collagen, Colony

stimulating factor (CSF), Complement factor 5a, Complement inhibitor, Complement

receptor 1, Factor IX, Factor VII, Factor VIII, Factor X, Fibrinogen, Fibronectin,

Glucocerebrosidase, Gonadotropin, Hedgehog proteins (e.g., Sonic, Indian, Desert),

30    Hemoglobin (for blood substitute; for radiosensitization), Hirudin, Human serum albumin,

Lactoferrin, Luciferase, Neurturin, Neutrophil inhibitory factor (NIF), Osteogenic protein,

Parathyroid hormone, Protein A, Protein G, Relaxin, Renin, Salmon calcitonin, Salmon growth hormone, Soluble complement receptor I, Soluble I-CAM 1, Soluble interleukin receptors (IL-1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15), Soluble TNF receptor, Somatomedin, Somatostatin, Somatotropin, Streptokinase, Superantigens, i.e.,

5      Staphylococcal enterotoxins (SEA, SEB, SEC1, SEC2, SEC3, SED, SEE), Toxic shock syndrome toxin (TSST-1), Exfoliating toxins A and B, Pyrogenic exotoxins A, B, and C, and M. arthritides mitogen, Superoxide dismutase, Thymosin alpha 1, Tissue plasminogen activator, Tumor necrosis factor beta (TNF beta), Tumor necrosis factor receptor (TNFR), Tumor necrosis factor-alpha (TNF alpha) and Urokinase. Many other known coding nucleic

10     acids, such as those in Genebank™, can be codon-altered and shuffled.

## GENES WITH CODON USAGE REDESIGNED AND CHEMICALLY SYNTHESIZED AS STARTING MATERIALS FOR GENE FAMILY SHUFFLING--EXPANDING THE DIVERSITY OF DNA SHUFFLING.

Because the genetic coding preference among organisms ranges from quite

15     similar to very different, homologous genes from different organisms can have significantly lower homology at the nucleic acid level than at the amino acid level. For example, genetic information for some bacterial species is high in GC content (up to 70%), while others have AT rich (>60%) codon usage. Thus, genes from different organisms may have, for example, 40-60% amino acid identity but only 25-35% nucleic acid identity. It is often desirable to

20     increase such levels of nucleic acid identity so as to enhance the ability of the homologous sequences to recombine, thereby increasing the efficiency of family shuffling using the methods of this invention. In other aspects, it is actually preferable to decrease the rate of recombination in a system, e.g., when using vectors it is sometimes desirable to decrease the rate of recombination between the vector and the host DNA, thereby increasing the safety of

25     the vector. The following examples address specific issues with regard to shuffling codon altered nucleic acids.

### Altering codon usage to increase homology

In one aspect, protein sequences of gene family members are reverse translated back into DNA sequences, for example by using one of the preferable codon usage

30     charts in any conventional DNA manipulation program (e.g. the Wisconsin Package™, SeqWeb, OMIGA, SeqApp, SeqPup, MacVector, DNA stryder, GeneWorks, etc.). The

choice of codon usage is often determined by the host in which the genes will be expressed. After maximizing the percentage of DNA sequence identity, the genes are chemically synthesized, *e.g.*, using a high throughput oligonucleotide synthesizer in, *e.g.* a 96-well format, optionally in conjunction with polymerase and/ or ligase gene synthesis methods.

5          In general, the DNA sequence similarity after such treatment will be at least as high as the amino acid similarity, but can be at least about 10% to 15% *higher* than the amino acid identity (in contrast to the situation for naturally occurring genes, which are ordinarily less well conserved than encoded polypeptides), based on the random frequency of sequence identity for any given codon. In most cases, the minimal requirement for amino acid identity

10        can be as low as about 35% while still retaining adequate nucleic acid homology for standard recombination methods (as discussed, *supra*, oligonucleotide-mediated recombination methods do not require high levels of similarity to achieve recombination). In some cases, however, the minimal amino acid identity can be even lower, *e.g.* if the conserved regions are clustered within the genes.

15        <u>Example: Shuffling codon-modified EPO</u>

The protein erythropoietin alpha, also known as EPO, Epogen, and Procrit is a hematopoietic hormone, providing a variety of benefits to patients suffering from anemia (a common symptom of, *e.g.*, AIDS). EPO is produced as a pharmaceutical, with sales of nearly 1 billion dollars world-wide. Accordingly, proteins with EPO-like activity (and

20        preferably superior activity) are of substantial commercial interest.

Figure 1 shows the sequence of a part of the monkey EPO gene, which is similar to the human EPO gene. Figure 2 shows an example of a codon altered EPO nucleic acid (or "wobble" EPO gene). In general, transversions rather than transition mutations are made where possible. The purpose of this strategy is to maximally disrupt hybridization of

25        the resulting gene with naturally occurring EPOs. Figure 3 shows an alignment of naturally occurring EPOs.

This strategy is further fine-tuned by applying standard rules of base pairing (e.g., elimination of G-C pairing and GC stacking) to maximize sequence disruption; in addition, conservative or non-conservative amino acid modifications can also be made (in

30        some cases, where multiple codon-altered nucleic acids are shuffled, it is desirable to make codon altered nucleic acids with non-overlapping non-conservative substitutions to permit

26

reversion to the wild-type amino acid during shuffling). The size of the sequence space for nucleic acids encoding EPO is large, at about $2.8 \times 10^{88}$ different sequences (there are about $10^{80}$ particles in the universe; thus, it is physically impossible to make all of the possible sequences encoding EPO). As indicated schematically in Figure 4, if one only considers the maximally divergent wobble genes (those that use the alternative types of codons for leucine, arginine, and serine), there is still a sequence space of $10^{38}$ sequences encoding EPO. The overall strategy is to synthesize a library of wobble genes, screen for expression and activity and DNA shuffle desirable genes as desired (e.g., by recursive processes).

It is of interest to further evolve codon-altered nucleic acids. Shuffling with other homologous genes from nature, designed genes (incorporating libraries of designed sequence variation), and genes containing mutations of interest are strategies for evolving any gene of interest. However, the codon altered nucleic acid may not be easily shuffled with these genes because of the sequence differences; or they may be undesirable for other reasons (e.g., the naturally occurring sequences may be proprietary, or include proprietary elements). These difficulties can be avoided by synthesizing codon-altered homologous nucleic acids which encode desired amino acid variations (e.g., those found in homologous genes), but which have a codon-set close to the nucleic acid(s) to be recombined (thereby permitting, e.g., hybridization during recombination).

For example, after identifying homologues of interest (e.g., those shown in Figure 3 for EPO), codon-altered nucleic acids encoding the same proteins are synthesized with a similar codon selection. Standard family shuffling is then practiced with the codon altered nucleic acids. This is shown schematically for EPO in Figure 5.

EPO wobble variants are screened for expression and then receptor binding assays are conducted in an ELISA format, using human EPOr-Fc fusions. Following selection of binding variants, activity is measured as thymidine incorporation in UT7-EPO (A human bone marrow cell line) cell proliferation assays. Cells are treated for 2-3 days with various concentration of EPO variants after which time they are incubated in the presence of 3-H thymidine for 4 hours and incorporation of thymidine is measured. See also, Erickson-miller et al. (1997) Blood 90:2421 (for the receptor binding assay), and Wen et al. (1994) J. Biol. Chem. 269:22839-22846 (for the thymidine incorporation assay).

Assays for selecting EPO can also be based, e.g., on the ability of EPO proteins to stimulate the growth of blood cell, e.g., in vitro or in vivo.

Example: Codon Shuffling G-CSF

Family shuffling can be used to breed diversity from genes into the libraries to be screened. Additionally, design heuristics such as randomization of hydrophobic core residues can be used to take advantage of the redundancy between primary structure and tertiary structure of proteins (i.e. many different primary structures encode proteins with very similar three dimensional structures).

Design heuristics are employed to create a sequence space of mutants that are predicted to be highly biased (relative to random mutagenesis) to encode proteins which preserve the original activity. Methods such as high throughput (HTP) screening and phage panning are used to identify members of the designed libraries that have the desired activity. DNA shuffling is used to breed this population of active clones in order to fine tune the mutants, thus allowing one to evolve variants with equivalent or superior function relative to the naturally occurring proteins.

Figures 6 and 7 show several mammalian homologues of G-CSF. Figure 8 shows the hydrophobic core residues of human G-CSF (blacked out). Figure 9 shows a strategy for evolving variants of human G-CSF that are highly divergent in sequence. First, three genes are synthesized (Genes 1, 2 and 3, Figure 8) which contain all of the mammalian homologue diversity of G-CSF. These genes are shuffled, phage panned against the G-CSF receptor, and HTP screened for biological function (receptor activation). Active clones are iteratively shuffled and screened if necessary to give evolved variants that rival or surpass the human gene in activity (on human cells).

Next, one evolves a variant that has a highly mutated hydrophobic core. This is schematically illustrated in Figure 8, and the specific strategy for performing the biological screening is schematically illustrated in Figure 9. it is expected that the best mutants obtained after screening hydrophobic core randomized libraries may be less active than wild type human G-CSF because it is difficult to initially optimize activity in such a procedure. Family shuffling is used to obtain optimized variants. This is done by synthesizing genes which contain mammalian homologue diversity at all but the hydrophobic core positions; but

28

they are synthesized in the context of an evolved, non-wild type hydrophobic core. Family shuffling is used to optimize around the new hydrophobic core.

This strategy works because there are functionally similar hydrophobic cores for wild type proteins that consist of largely different amino acids than the wild type protein.

5 This understanding is supported by recent experiments in model systems. For example, 53% of randomized sequences for three residues in the hydrophobic core of lambda repressor are folded and biologically active (Lim and Sauer (1991) J. Mol. Biol. 219:359-376). Protein design by patterning of polar and non-polar amino acids, where 24 residues in the hydrophobic core of a 4-helix bundle protein were randomized by Kamtekar et al. (1993)

10 Science 262:1321. Folded, alpha helical proteins were recovered from about 1% of the clones. Desjarlais and Handel (1995) Current Opinion in Biotechnology 6:460-466 showed a mutant of Rop, another 4-helix bundle protein, where four hydrophobic core residues have been randomized and active mutants have been obtained. Axe et al. (1996) PNAS 95:5590-5594 showed that randomizing 13 hydrophobic core residues in the enzyme barnase resulted

15 in 23% of the clones in the library retaining biological activity. Gassener et al. (1996) PNAS 93:12155-12158 describe a mutant of T4 lysozyme where 10 residues in the hydrophobic core are replaced with Met. This is taken as evidence that the hydrophobic core of this protein is very tolerant to substitution. Taken together, this experimental evidence on model systems shows that the hydrophobic cores of many proteins can be replaced with other

20 hydrophobic residues that pack in a similar fashion to give an active protein. This degeneracy is exploited to evolve novel forms of natural and codon-altered genes.

A related approach is to search the protein databases for a protein that has a similar activity to a protein that on wishes to evolve. Denesyuk et al. (1996) J. Theor. Biol shows the results of such a search for G-CSF. LIF is a very similarly folded protein. One

25 can use LIF as a 'scaffold' on which to place residues of G-CSF that are required for activity. Given LIF with a G-CSF "toupee," one would family shuffle the LIF scaffold so as to obtain a variant in which the toupee is displayed in a fully biologically active form.

Another approach is to use computational methods to create families of variants that are predicted to be functional. Dahiyat and Mayo Science recently described

30 computer methods that are used to design proteins. Proteins are simulated on the computer, often with the aid of genetic algorithms, and a subset that are deemed 'fit' are actually

synthesized and 'analyzed'. These computational methods are becoming increasingly powerful. They would be useful to, for example, predict a family of mutations on the surface of a protein that would not destroy function. DNA shuffling can be used to optimized active clones obtained by design. Taking the example of G-CSF, one could use computational

5    methods in combination with all structure function data (for example alanine scan data for G-CSF reported recently by Reidhaar-Olsen in Biochemistry) to design a family of putatively functional variants. One could, for example, design the family to have minimal DNA identity to the wild type gene given the design constraints. This library is synthesized, put through biological screens and/or selections (i.e. panning against the G-CSF receptor), and active

10   variants are obtained. DNA shuffling is then used to evolve these active variants to have the desired level of function.

A G-CSF proteins are displayed on phage and screened for binding to human G-CSF receptor in an ELISA format. Variants that bind receptor are selected in a high throughput screen for receptor activation. This cell based assay measures receptor activation

15   via a reporter gene (such as luciferase) activated by a G-CSF responsive construct containing STAT binding elements. Cells (such as HepG2) are transformed with a G-CSF responsive reporter plasmid and treated with the codon shuffled G-CSF variant for 2.5 hours. Cells are then lysed and luciferase activity measured. *See also*, Tian et al (1998) Science 281:257-259.

### Example: Codon Shuffling Alkaline Phosphatase

20   Alkaline phosphatase is a widely used reporter enzyme for ELISA assays, protein fusion assays, and in a secreted form as a reporter gene for mammalian cells. A more active form of the enzyme is desirable.

A codon altered form of alkaline phosphatase was generated by PCR assembly using the oligos set forth in Figure 10. A map of the oligos is set forth in Figure 11. The

25   procedure used was essentially identical to that taught in Stemmer et al. (1994) Gene 164:49-57. In brief, the oligos were mixed 1:1 at a variety of dilutions and PCR assembled by performing e.g., 25-60 cycles of PCR at e.g., 94 °C (60 sec.), 94 °C (30 sec.), 50 °C (30 sec.), 72 °C (30 sec). Assembly of the BIAP gene was conducted in a circular format and gene fragments were purified. ~100,000 colonies were screened on LB/am plates (~1/10 are wt

30   plasmid). About 1/10 showed a bluer color than background. Plasmid DNA showed a correct insertion.

In general, petri-dish screening using the typical colorimetric assay for phosphatase activity can be used for screening. This has the advantage or being simple, high throughput, and semi quantitative. Microtiter plate screening, also preferred, is colorimetric, and quantitative, although additional instrumentation can be required for implementation.

5          Example: Codon Shuffling to Reduce Competent Virus Production from
           Vectors and to Generate Attenuated Viruses as Immunogenic Compositions
           and Vaccines
           Cells can be stably transduced with a number of viral vectors including those derived from retroviruses, pox viruses, adenoviruses (Ads), herpes viruses and parvoviruses.

10   Common viral vectors include those derived from murine leukemia viruses (MuLV), gibbon ape leukemia viruses (GaLV), human immuno deficiency viruses (HIV), adenoviruses, adeno associated viruses (AAVs), Epstein Barr viruses, canarypox viruses, cowpox viruses, and vaccinia viruses. Viral vectors based upon retroviruses, adeno-associated viruses, herpes viruses and adenoviruses are all used as gene therapy vectors for the introduction of

15   therapeutic nucleic acids into the cells of an organism by *ex vivo* and *in vivo* methods.

When using viral vectors, packaging cells are commonly used to prepare virions used to transduce target cells. In these vectors, trans-active genes are rendered inactive and "rescued" by trans-complementation to provide a packaged vector. This form of trans complementation is provided by co-infection of a packaging cell with a virus or vector

20   which supplies functions missing from a particular gene therapy vector in trans, or by using a cell line (e.g., 293 cells) which have viral components integrated into the genome of the packaging cell. For instance, cells transduced with HIV or murine retroviral proviral sequences which lack the nucleic acid packaging site produce retroviral trans active components, but do not specifically incorporate the retroviral nucleic acids into the capsids

25   produced, and therefore produce little or no live virus.

If these transduced "packaging" cells are subsequently transduced with a vector nucleic acid which lacks coding sequences for retroviral trans active functions, but includes a packaging signal, the vector nucleic acid is packaged into an infective virion. A number of packaging cell lines useful for MoMLV-based vectors are known in the art, such

30   as PA317 (ATCC CRL 9078) which expresses MoMLV core and envelope proteins see, Miller et al. J. Virol. 65:2220-2224 (1991). Carrol et al. (1994) Journal of virology 68(9):6047-6051 describe the construction of packaging cell lines for HIV viruses.

31

Reciprocal complementation of defective HIV molecular clones is described, e.g., in Lori et al. (1992) Journal of Virology 66(9) 5553-5560.

Functions of viral replication not supplied by trans-complementation which are necessary for replication of the vector are present in the vector. In HIV, this typically

5    includes, e.g., the TAR sequence, the sequences necessary for HIV packaging, the RRE sequence if the instability elements of the p17 gene of gag is included, and sequences encoding the polypurine tract. HIV sequences that contain these functions include a portion of the 5' long terminal repeat (LTR) and sequences downstream of the 5' LTR responsible for efficient packaging, i.e., through the major splice donor site ("MSD"), and the polypurine

10   tract upstream of the 3' LTR through the U3R section of the 3' LTR. The packaging site (psi site or ψ site) is partially located adjacent to the 5' LTR, primarily between the MSD site and the gag initiator codon (AUG) in the leader sequence. See, Garzino-Demo et al. (1995) Hum. Gene Ther. 6(2): 177-184. For a general description of the structural elements of the HIV genome, see, Holmes et al. PCT/EP92/02787.

15   Another common vector is based upon adenovirus. Typically, vectors which include the adenovirus ITRs (Gingeras et al. (1982) J. Biol. Chem. 257:13475-13491) are packaged in, e.g., 293 cells, which provide many of the components necessary for vector packaging.

Adeno-associated viruses (AAVs) utilize helper viruses such as adenovirus or

20   herpes virus to achieve productive infection. In the absence of helper virus functions, AAV integrates (site-specifically) into a host cell's genome, but the integrated AAV genome has no pathogenic effect. The integration step allows the AAV genome to remain genetically intact until the host is exposed to the appropriate environmental conditions (e.g., a lytic helper virus), whereupon it re-enters the lytic life-cycle. Samulski (1993) Current Opinion in

25   Genetic and Development 3:74-80 and the references cited therein provides an overview of the AAV life cycle. For a general review of AAVs and of the adenovirus or herpes helper functions see, Berns and Bohensky (1987) Advanced in Virus Research, Academic Press., 32:243-306. The genome of AAV is described in Laughlin et al. (1983) Gene, 23:65-73. Expression of AAV is described in Beaton et al. (1989) J. Virol., 63:4450-4454. In general,

30   the packaging sites for all parvoviruses, including B 19 and AAV are located in the viral ITRs. Recombinant AAV vectors (rAAV vectors) deliver foreign nucleic acids to a wide

32

range of mammalian cells (Hermonat & Muzycka (1984) *Proc Natl Acad Sci USA* 81:6466-6470; Tratschin *et al.* (1985) *Mol Cell Biol* 5:3251-3260), integrate into the host chromosome (McLaughlin *et al.* (1988) *J Virol* 62: 1963-1973), and show stable expression of the transgene in cell and animal models (Flotte *et al.* (1993) *Proc Natl Acad Sci USA* 90:10613-

5    10617). rAAV vectors are able to infect non-dividing cells (Podsakoff *et al.* (1994) *J Virol* 68:5656-66; Flotte *et al.* (1994) *Am. J. Respir. Cell Mol. Biol.* 11:517-521). Further advantages of rAAV vectors include the lack of an intrinsic strong promoter, thus avoiding possible activation of downstream cellular sequences, and the vector s naked icosohedral capsid structure, which renders the vectors stable and easy to concentrate by common

10   laboratory techniques.

       One problem with previously existing vector packaging strategies is that vectors to be packaged can recombine with nucleic acids providing packaging functions in trans, producing a replication-competent virus. This can be a problem both when vectors are produced for therapeutic applications (e.g., in gene therapy) and during production of

15   encoded components in vitro. The present invention provides a way of reducing or eliminating recombination between nucleic acids encoding trans-active components and vector nucleic acids encoding packaging sites.

       In particular, nucleic acid subsequences of a vector which are adjacent to modified or deleted elements provided in trans, are codon modified to eliminate hybridization

20   to wild-type sequences. Because these sequences do not hybridize, they cannot recombine with nucleic acids producing trans-active components. One additional advantage of this approach is that the vectors also cannot recombine with live viruses, e.g., in a human body which is infected with a virus that packages vector elements. As noted above, two types of gene therapy vectors are those based upon retroviruses (which can be packaged by, e.g., HIV-

25   1) and adenoviruses (which can be packaged by adenovirus).

       Alternatively, the nucleic acids encoding trans-active components can be codon modified so that they do not hybridize to wild-type sequences. This also prevents recombination with vectors having wild-type sequences, preventing recombination and formation of replication competent viruses.

33

After codon modification, vectors or trans active nucleic acids can be shuffled as described *supra*, and screened for the ability to package nucleic acids, or to be packaged, as appropriate.

It will be appreciated that codon modification of viral sequences has an

5    additional use as well. Codon alteration of viral sequences can result in attenuation of the virus, e.g., due to modification of regulatory sequences, alterations in mRNA secondary structure, inefficient translation due to rare codon use, and the like. Such "codon attenuated" viruses have a significant advantage over existing attenuated viruses (which are typically generated by serial passage in cells other than the normal host type for the virus). In

10   particular, codon attenuated viruses can encode a wild-type set of proteins, making them ideal as immunogenic compositions to generate antibodies, or to use as vaccines. Viral proteins can also be used in various diagnostic assays. For example, the standard diagnostic test for HIV infection in current use tests for the presence of anti-HIV antibodies in blood by probing with viral proteins.

15   Example: Codon Usage Libraries to evolve Functional Variants with reduced
     Recombination With Natural Gene Sequences--Adenovirus
     Adenovirus is a common vector used, e.g., for gene therapy. The virus is typically modified to make it replication deficient. This can be achieved e.g., by deleting the E1 and E4 genes. The functions of E1 and E4 can be supplied by trans complementation

20   when E1 and E4 deleted vectors are grown in the ubiquitous human embryonic kidney cell line 293, which has uncharacterized adenovirus fragments incorporated into their genome that supply the missing functions in trans. The replication defective adenoviral vectors recombine at a low, but clinically significant frequency, resulting in replication competent adenovirus contamination of vector preparations. Because adenovirus has detrimental effects

25   on health, this is a significant problem for application of adenovirus-based gene therapy vectors.

In the present invention, a codon usage library encompassing several hundred bases to several kilobases of sequence flanking the adenovirus E1 and E4 genes are made. The library is designed to enforce a high degree of divergence from the natural adenoviral

30   consensus sequence, while at the same time incorporating a large degree of degeneracy in the codons to allow for a large space of sequence diversity to be searched. The design principle

is to obtain mutants that encode the same or similar protein sequence, but with many mismatches to the wild-type E1 and E4 sequences found in the 293 genome. These mismatches strongly reduce the frequency of unwanted recombination with the trans complementary genes. Consequently, engineered adenoviral vectors, or adenovirus helper

5     vectors which package adenoviral sequences which include packaging sequences (adenoviral or adeno-associated viral ITRs) in trans have reduced levels of recombination. This provides for a lower rate of competent adenovirus production, making culture and production of such vectors safer.

### Evolution Impaired Viruses Created by Massive Codon Usage Alteration As a General Approach to Vaccines--HIV

10     HIV-1 and HIV-2 are genetically related, antigenically cross reactive, and share a common cellular receptor (CD4). *See*, Rosenburg and Fauci (1993) in *Fundamental Immunology, Third Edition* Paul (ed) Raven Press, Ltd., New York (Rosenburg and Fauci 1) and the references therein for an overview of HIV infection. HIV-1 infection is epidemic

15     world wide, causing a variety of immune system-failure related phenomena commonly termed acquired immune deficiency syndrome (AIDS). HIV type 2 (HIV-2) has been isolated from both healthy individuals and patients with AIDS-like illnesses (Andreasson, *et al.* (1993) *Aids* 7, 989-93; Clavel, *et al.* (1986) *Nature*, 324, 691-695; Gao, *et al.* (1992) *Nature* 358, 495-9; Harrison, *et al.* (1991) *Journal of Acquired Immune Deficiency*

20     *Syndromes* 4, 1155-60; Kanki, *et al.* (1992) *American Journal of Epidemiology* 136, 895-907; Kanki, *et al.* (1991) *Aids Clinical Review* 1991, 17-38; Romieu, *et al.* (1990) *Journal of Acquired Immune Deficiency Syndromes* 3, 220-30; Naucler, *et al.* (1993) *International Journal of STD and Aids* 4, 217-21; Naucler, *et al.* (1991) *Aids* 5, 301-4). Although HIV-2 AIDS cases have been identified principally from West Africa, sporadic HIV-2 related AIDS

25     cases have also been reported in the United States (O'Brien, *et al.* (1991) *Aids* 5, 85-8) and elsewhere. HIV-2 will likely become endemic in other regions over time, following routes of transmission similar to HIV-1 (Harrison, *et al.* (1991) *Journal of Acquired Immune Deficiency Syndromes* 4, 1155-60; Kanki, *et al.* (1992) *American Journal of Epidemiology* 136, 895-907; Romieu, *et al.* (1990) *Journal of Acquired Immune Deficiency Syndromes* 3,

30     220-30). Epidemiological studies suggest that HIV-2 produces human disease with lesser penetrance than HIV-1, and exhibits a considerably longer period of clinical latency (at least

25 years, and possibly longer, as opposed to less than a decade for HIV-1; *see*, Kanki, *et al.*
(1991) *Aids Clinical Review* 1991, 17-38; Romieu, *et al.* (1990) *Journal of Acquired Immune*
*Deficiency Syndromes* 3, 220-30, and Travers *et al.* (1995) *Science* 268: 1612-1615).

The ability of HIV virus populations to rapidly point mutate to avoid the

5   immune response poses a special challenge for vaccine design. While the immune system
has responded to viruses in a gradual and co-evolutionary manner, the present invention
provides a general approach that provides for massively faster evolution to produce new
vaccines to stimulate more effective immune responses.

For example, during the incubation period for HIV infection, which lasts for

10  several years, low titers of HIV can result from high HIV replication rates in conjunction
with efficient viral clearance by the immune system. In response to these selective forces,
virus mutations are selected which reduce recognition and neutralization by the immune
system's B and T-cell responses. *See*, Lukashov et al. (1995) J. Virol. 69:6911-6916. During
the long incubation time, these mutations accumulate and eventually overwhelm the immune

15  system's defenses. *See*, Ho et al. (1995) Nature.

Live attenuated vaccines, typically produced by prolonged growth of human
viruses in animal cells, have proven useful as vaccines for several diseases, including mumps,
rubella and measles. Attenuation involves the slow accumulation of many mutations
throughout the viral genome during the course of adapting to growth in the animal cells.

20  When used to vaccinate humans, the attenuated virus grows only weakly and elicits a
complex immune response which the virus is unable to avoid. The mutations in the
attenuated virus could, in principle, revert in the same stepwise fashion that it underwent to
grow in culture.

The risk of reversion is highest in viruses with a high mutation rate such as

25  HIV-1, which makes this strategy dangerous under current techniques for vaccine
development. It is worth noting, however, that protective effects against HIV-1 are observed
following infection with the related HIV-2 virus, which is much less pathogenic than HIV-1.
Thus, protective effects against HIV can be achieved with live vaccines.

To reduce the risk of reversion, a large number of mutations need to

30  accumulate in the virus. However, if too many mutations are present, the immune system in

effect recognizes the attenuated virus, but not the virus against which a protective effect is sought.

As provided herein, immunogenic compositions such as vaccines are created which contain a large number of silent substitutions. In contrast to existing attenuated

5    viruses, such viruses have native protein sequences and elicit essentially the same immune responses as the corresponding wild-type virus (typically one or a few additional disabling mutations can also be incorporated). Codon alteration results in two effects that both increase the potential of the vaccine.

First, like standard attenuated viruses, the growth of codon-altered viruses is

10   attenuated, due to the effect of the codon alterations on translation, regulatory sequences, mRNA folding, packaging, and the like. For example, regulation of HIV-1 envelope expression has been observed as a result of codon usage. *See,* Haas et al. (1989) Current Biology 6(3):315-324.

Second, codon alteration results in impairment of virus evolution. As

15   discussed above, modification of the codons alters the mutational escape spectrum of the virus, upsetting the evolutionary selection for specific codons.

The six codon amino acids are the best targets for codon alteration. Serine, arginine and leucine each have one group of four codons, plus two codons in an unrelated group. *See,* Figure 12. Switching all of the serine codons from AGY to TCX and vice versa,

20   yields proteins with unaltered amino acid sequences. *See also,* Figure 13. However, these codon groups differ significantly in the spectrum of the amino acids that they yield upon point mutation. Of all possible point mutations of one codon for serine (TCA) 78% result in a different amino acid compared to point mutations obtained for the AGT codon for serine. *See,* Figure 13. A virus with hundreds of codon alterations is in, statistically, a very different

25   mutational space, able to access a totally different mutation spectrum, or "cloud," compared to the wild-type virus. The overall strategy for producing an evolution-defective virus is additionally set forth in Figures 14, 15 and 17. Figure 16, panels A-C show results of single mutations of different codons for ser, arg, and leu.

Point mutation is critical for viruses such as HIV-1 to stay ahead of the host

30   immune system. The amino acid mutations that are required for virus escape are likely not random. Wild type codon usage has evolved to allow optimal immune system evasion. The

wild type codon usage is likely to favor mutations that represent alterations that avoid the host immune system, without detrimentally affecting the protein(s) encoded. While complex, this natural pattern of amino acid sequence change of the natural virus in response to the host system is non-random and weakly predictable. *See also*, Seiller-Moiseiwitsch et al. (1994)

5     Annu. Rev. Genet. 28:559-596.

Changing all of the codons for ser, arg and leu in the 875 aa envelope polyprotein of HIV-1 (e.g., strain MN) would affect 187 codons (22%) resulting in 561 mutations. *See*, e.g., Figure 18, panels A-D. If all of the HIV proteins were altered, the number of mutations would be more than three-fold higher. The construction of such codon

10     modified viruses is simplified by recent advances in the synthesis of long DNA sequences, which enable the assembly of a plasmid of average size from 40 mer oligos in a single step with about 75% efficiency. *See*, Stemmer (1994) Nature 370 389-391. *See also*, Figure 19 for a list of oligos in one application for synthesis of HIV Env. While the synthesis of the envelope gene is sufficient, synthesis of the whole HIV genome from oligos can be

15     performed by this method.

In practice, a preferred balance of attenuation and evolution impairment is obtained by DNA shuffling (e.g., Stemmer et al. (1995) Gene 164:49-53), e.g., of the wild-type and codon altered sequences, followed by selection of the resulting library of viruses that retain moderate growth despite many codon alterations.

20     While attenuation that can be obtained by this approach may be sufficient for obtaining a vaccine for most viruses, for HIV-1, the evolution impairment is more important, due to the high mutation rate of the virus. Live vaccines are used only if they elicit an immune response which is complex and strong enough to prevent infection of the wild-type virus. Live virus vaccines are typically more protective than single protein vaccines because

25     it is harder to out-mutate T and B-cell responses to a larger number of epitopes. The weak growth of the live virus vaccine results in a larger antigenic dose and point mutation is increases the complexity of the immune response. To evaluate vaccine competence, vaccine potential is evaluated in Macaques (*M. nemestrina*) or chimpanzees using SIV variants that are known to cause AIDS. Sequence for an example SIV, SIVsmm, is found at Gene Bank

30     Accession No. x14307. This virus is closely related to HIV-2. *See*, Hirsch (1989) Nature 339: 389-392. In general, many complete sequences for HIVs, SIVs and many other viruses

are found in well known sequence repositories, including GenBank, EMBL, DDBJ and the NCBI. Well characterized HIV clones include: HIV-1NL43, HIV-1SF2, HIV-1BRU and HIV-1MN. For an introduction to the genetic variability of HIV, *see*, Seillier-Moiseiwitsch et al. (1994) Annu. Rev. Genet. 28:559-96 and the references cited therein.

5          Several HIV-2 isolates, including three molecular clones of HIV-2 (HIV-2$_{ROD}$, HIV-2$_{SBL-ISY}$, and HIV-2$_{UC1}$), have also been reported to infect macaques (*M. mulatta* and *M. nemestrina*) or baboons (Franchini, *et al.* (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 2433-2437; Barnett, *et al.* (1993) *Journal of Virology* 67, 1006-14; Boeri, *et al.* (1992) *Journal of Virology* 66, 4546-50; Castro, *et al.* (1991) *Virology* 184, 219-26; Franchini, *et al.* (1990)
10        *Journal of Virology* 64, 4462-7; Putkonen, *et al.* (1990) *Aids* 4, 783-9; Putkonen, *et al.* (1991) *Nature* 352, 436-8). As human pathogens capable of infection of small primates, HIV-2 molecular clones provide attractive models for studies of AIDS pathogenesis, and for drug and vaccine development against HIV-1 and HIV-2.

          Recently, HIV-2 was suggested as a possible vaccine candidate against the
15        more virulent HIV-1 due to its long asymptomatic latency period, and its ability to protect against infection by HIV-1 (*see*, Travers *et al.* (1995) *Science* 268: 1612-1615 and related commentary by Cohen *et al* (1995) *Science* 268: 1566). In the nine-year study by Travers *et al.* (*id*) of West African prostitutes infected with HIV-2, it was determined that infection with HIV-2 caused a 70% reduction in infection by HIV-1. Thus, codon altered HIV-2 viruses can
20        also be used as a live vaccine, against both HIV-2 and HIV-1. Furthermore, because the natural pathogenicity of HIV-2 is less than HIV-1, it is, in addition to HIV-1, a preferred virus for modification.

## FORMATS FOR SEQUENCE RECOMBINATION

          The methods of the invention entail performing recombination ("shuffling")
25        and screening or selection to "evolve" individual genes, whole plasmids or viruses, multigene clusters, or even whole genomes (Stemmer (1995) *Bio/Technology* 13:549-553). Reiterative cycles of recombination and screening/selection can be performed to further evolve the nucleic acids of interest. Such techniques do not require the extensive analysis and computation required by conventional methods for polypeptide engineering. Shuffling
30        allows the recombination of large numbers of mutations in a minimum number of selection cycles, in contrast to natural pair-wise recombination events (e.g., as occur during sexual

replication). Thus, the sequence recombination techniques described herein provide particular advantages in that they provide recombination between mutations in any or all of these, thereby providing a very fast way of exploring the manner in which different combinations of mutations can affect a desired result. In some instances, however, structural

5    and/or functional information is available which, although not required for sequence recombination, provides opportunities for modification of the technique.

Exemplary formats and examples for sequence recombination, referred to, e.g., as "DNA shuffling," "fast forced evolution," or "molecular breeding," have been described by the present inventors and co-workers in the following patents and patent

10   applications: US Patent No. 5,605,793; PCT Application WO 95/22625 (Serial No. PCT/US95/02126), filed February 17, 1995; US Serial No. 08/425,684, filed April 18, 1995; US Serial No. 08/621,430, filed March 25, 1996; PCT Application WO 97/20078 (Serial No. PCT/US96/05480), filed April 18, 1996; PCT Application WO 97/35966, filed March 20, 1997; US Serial No. 08/675,502, filed July 3, 1996; US Serial No. 08/721, 824, filed

15   September 27, 1996; PCT Application WO 98/13487, filed September 26, 1997; "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination" Attorney Docket No. 018097-020720US filed July 15, 1998 by del Cardayre et al. (PCT/US99/15972, filed 07/15/1999); Stemmer, *Science* 270:1510 (1995); Stemmer *et al.*, *Gene* 164:49-53 (1995); Stemmer, *Bio/Technology* 13:549-553 (1995); Stemmer, *Proc. Natl. Acad. Sci. U.S.A.*

20   91:10747-10751 (1994); Stemmer, *Nature* 370:389-391 (1994); Crameri *et al.*, *Nature Medicine* 2(1):1-3 (1996); and Crameri *et al.*, *Nature Biotechnology* 14:315-319 (1996), each of which is incorporated by reference in its entirety for all purposes.

The recombination procedure starts with at least two substrates that generally show substantial sequence identity to each other (*e.g.*, at least about 30%, 50%, 70%, 80% or

25   90% or more sequence identity), but differ from each other at certain positions. For example, at least one codon altered nucleic acid is recombined with one or more additional nucleic acid (the additional nucleic acid can also be a codon altered nucleic acid) herein. The difference between nucleic acids to be recombined can be any type of mutation, for example, substitutions, insertions and deletions. Often, different segments differ from each other in

30   about 5-20 positions. For recombination to generate increased diversity relative to the starting materials, the starting materials must differ from each other in at least two nucleotide

40

positions. That is, if there are only two substrates, there should be at least two divergent positions. If there are three substrates, for example, one substrate can differ from the second at a single position, and the second can differ from the third at a different single position.
The starting DNA segments can be natural variants of each other, for example, allelic or
5    species variants. More typically, they will be codon altered nucleic acids derived from one or more homologous nucleic acid sequence. The segments can also be from nonallelic genes showing some degree of structural and usually functional relatedness (*e.g.*, codon altered nucleic acids derived from different, but homologous, genes within a superfamily). The starting DNA segments can also be induced variants of each other. For example, one DNA
10   segment can be produced by error-prone PCR replication of the other, or by substitution of a mutagenic cassette. Induced mutants can also be prepared by propagating one (or both) of the segments in a mutagenic strain. In these situations, strictly speaking, the second DNA segment is not a single segment but a large family of related segments. The different segments forming the starting materials are often the same length or substantially the same
15   length. However, this need not be the case; for example; one segment can be a subsequence of another. The segments can be present as part of larger molecules, such as vectors, or can be in isolated form.

The starting DNA segments are recombined by any of the sequence recombination formats provided herein to generate a diverse library of recombinant DNA
20   segments. Such a library can vary widely in size from having fewer than 10 to more than $10^5$, $10^9$, $10^{12}$, $10^{15}$, $10^{20}$ or even more members. In some embodiments, the starting segments and the recombinant libraries generated will include essentially full-length coding sequences and any essential regulatory sequences, such as a promoter and polyadenylation sequence, required for expression. In other embodiments, the recombinant DNA segments in
25   the library can be inserted into a common vector providing sequences necessary for expression before performing screening/selection.

### Use of Restriction Enzyme Sites to Recombine Mutations
In some situations it is advantageous to use restriction enzyme sites in nucleic acids to direct the recombination of mutations in a nucleic acid sequence of interest. These
30   techniques are particularly preferred in the evolution of fragments that cannot readily be shuffled by other existing methods due to the presence of repeated DNA or other problematic

primary sequence motifs. These situations also include recombination formats in which it is preferred to retain certain sequences unmutated. The use of restriction enzyme sites is also preferred for shuffling large fragments (typically greater than 10 kb), such as gene clusters that cannot be readily shuffled and "PCR-amplified" because of their size. Although

5    fragments up to 50 kb have been reported to be amplified by PCR (Barnes, *Proc. Natl. Acad. Sci. U.S.A.* 91:2216-2220 (1994)), it can be problematic for fragments over 10 kb, and thus alternative methods for shuffling in the range of 10 - 50 kb and beyond are preferred. Preferably, the restriction endonucleases used are of the Class II type (Sambrook, Ausubel and Berger, *supra*) and of these, preferably those which generate nonpalindromic sticky end

10    overhangs such as Alwn I, Sfi I or BstX1. These enzymes generate nonpalindromic ends that allow for efficient ordered reassembly with DNA ligase. Typically, restriction enzyme (or endonuclease) sites are identified by conventional restriction enzyme mapping techniques (Sambrook, Ausubel, and Berger, *supra.*), by analysis of sequence information for that gene, or by introduction of desired restriction sites into a nucleic acid sequence by synthesis (*i.e.* by

15    incorporation of silent mutations). For example, one or more codon-altered nucleic acid can be recombined at restriction sites, e.g., with one or more nucleic acid of interest (including, e.g. a gene or gene cluster to be modified by recombination with the codon-altered nucleic acid).

        The DNA substrate molecules to be digested can either be from *in vivo*

20    replicated DNA, such as a plasmid preparation, or from synthetic or e.g., PCR amplified nucleic acid fragments harboring the restriction enzyme recognition sites of interest, preferably near the ends of the fragment. Typically, at least two variants of a gene of interest, each having one or more mutations, and at least one of which incorporating codon-modifications, are digested with at least one restriction enzyme determined to cut within the

25    nucleic acid sequence of interest. The restriction fragments are then joined with DNA ligase to generate full length genes having shuffled regions. The number of regions shuffled will depend on the number of cuts within the nucleic acid sequence of interest. The shuffled molecules can be introduced into cells as described above and screened or selected for a desired property as described herein. Nucleic acid can then be isolated from pools (libraries),

30    or clones having desired properties and subjected to the same procedure until a desired degree of improvement is obtained.

In some embodiments, at least one DNA substrate molecule or fragment thereof is isolated and subjected to mutagenesis. In some embodiments, the pool or library of religated restriction fragments are subjected to mutagenesis or additional recombination protocols before the digestion-ligation process is repeated. "Mutagenesis" as used herein

5      comprises such techniques known in the art as PCR mutagenesis, oligonucleotide-directed mutagenesis, site-directed mutagenesis, etc., and recursive sequence recombination by any of the techniques described herein.

Reassembly PCR

A further technique for recombining mutations in a nucleic acid sequence

10     utilizes "reassembly PCR." This method can be used to assemble multiple segments that have been separately evolved into a full length nucleic acid template such as a gene. This technique is performed when a pool of advantageous mutants is known from previous work or has been identified by screening mutants that may have been created by any mutagenesis technique known in the art, such as PCR mutagenesis, cassette mutagenesis, doped oligo

15     mutagenesis, chemical mutagenesis, or propagation of the DNA template *in vivo* in mutator strains. Boundaries defining segments of a nucleic acid sequence of interest preferably lie in intergenic regions, introns, or areas of a gene not likely to have mutations of interest. Preferably, oligonucleotide primers (oligos) are synthesized for PCR amplification of segments of the nucleic acid sequence of interest, such that the sequences of the

20     oligonucleotides overlap the junctions of two segments. The overlap region is typically about 10 to 100 nucleotides in length.

Each of the segments is amplified with a set of such primers. The PCR products are then "reassembled" according to assembly protocols such as those discussed herein to assemble randomly fragmented genes. In brief, in an assembly protocol the PCR

25     products are first purified away from the primers, by, for example, gel electrophoresis or size exclusion chromatography. Purified products are mixed together and subjected to about 1-10 cycles of denaturing, reannealing, and extension in the presence of polymerase and deoxynucleoside triphosphates (dNTPs) and appropriate buffer salts in the absence of additional primers ("self-priming"). Subsequent PCR with primers flanking the gene are used

30     to amplify the yield of the fully reassembled and shuffled genes. In some embodiments, the resulting reassembled genes are subjected to mutagenesis before the process is repeated.

In the present invention, oligos such as PCR primers can include codon modifications as compared to a starting sequence. In addition, oligonucleotides can form the basis for PCR concatemerization reactions in which overlapping hybridized oligonucleotides are extended in one or more PCR amplification cycles. In this embodiment, a template

5  nucleic acid is not required (although a template or fragments thereof can be added to the amplification mixture, which can aid in the eventual reassembly of a full-length gene). Further details regarding oligonucleotide gene reassembly methods are found, e.g., in Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed February 5, 1999, USSN 60/118,813 and Crameri et al. "OLIGONUCLEOTIDE

10  MEDIATED NUCLEIC ACID RECOMBINATION" filed June 24, 1999, USSN 60/141,049.

In a further embodiment, the PCR primers for amplification of segments of a nucleic acid sequence of interest are used to introduce variation into the gene of interest as follows. Mutations at sites of interest in a nucleic acid sequence are identified by screening or selection, by sequencing homologues of the nucleic acid sequence, and so on.

15  Oligonucleotide PCR primers are synthesized which encode wild type or mutant information at sites of interest. These primers are then used in PCR mutagenesis to generate libraries of full length genes encoding permutations of wild type and mutant information at the designated positions. This technique is typically advantageous in cases where the screening or selection process is expensive, cumbersome, or impractical relative to the cost of

20  sequencing the genes of mutants of interest and synthesizing mutagenic oligonucleotides.

#### Site Directed Mutagenesis (SDM) with Oligonucleotides Encoding Homologue Mutations Followed by Shuffling

In some embodiments of the invention, sequence information from one or more substrate sequences is added to a given "parental" sequence of interest, with subsequent

25  recombination between rounds of screening or selection. Typically, this is done with site-directed mutagenesis performed by techniques well known in the art (e.g., Berger, Ausubel and Sambrook, *supra.*) with one substrate as a template and oligonucleotides encoding single or multiple mutations from other substrate sequences, *e.g.* homologous genes. After screening or selection for an improved phenotype of interest, the selected recombinant(s) can

30  be further evolved using recursive techniques. After screening or selection, site-directed mutagenesis can be done again with another collection of oligonucleotides encoding

44

homologue mutations, and the above process repeated until the desired properties are
obtained.

When the difference between two homologues is one or more single point
mutations in a codon, degenerate oligonucleotides can be used that encode the sequences in
5    both homologues. One oligonucleotide can include many such degenerate codons and still
allow one to exhaustively search all permutations over that block of sequence.

When the homologue sequence space is very large, it can be advantageous to
restrict the search to certain variants. Thus, for example, computer modeling tools (Lathrop
*et al.* (1996) *J. Mol. Biol.*, 255: 641-665) can be used to model each homologue mutation
10   onto the target protein and discard any mutations that are predicted to grossly disrupt
structure and function. In silico genetic algorithm operations for generating and predicting
mutational events are found in Selifonov and Stemmer "METHODS FOR MAKING
CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED
CHARACTERISTICS" filed 02/05/1999, USSN 60/118854.

15                      Oligonucleotide and in silico shuffling formats
As mentioned above, at least two additional related formats are useful in the
practice of the present invention. The first, referred to as "in silico" shuffling utilizes
computer algorithms to perform "virtual" shuffling using genetic operators in a computer. As
applied to the present invention, codon altered gene sequence strings are recombined in a
20   computer system and desirable products are made, e.g., by reassembly PCR or ligation of
synthetic oligonucleotides. In silico shuffling is described in detail in Selifonov and Stemmer
in "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES &
POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" filed 02/05/1999, USSN
60/118854. In brief, genetic operators (algorithms which represent given genetic events such
25   as point mutations, recombination of two strands of homologous nucleic acids, etc.) are used
to model recombinational or mutational events which can occur in one or more nucleic acid,
e.g., by aligning nucleic acid sequence strings (using standard alignment software, or by
manual inspection and alignment) and predicting recombinational outcomes. The predicted
recombinational outcomes are used to produce corresponding molecules, e.g., by
30   oligonucleotide synthesis and reassembly PCR.

The second useful format is referred to as "oligonucleotide mediated shuffling" in which oligonucleotides corresponding to a family of related homologous nucleic acids (e.g., as applied to the present invention, codon modified synthetic homologous variants of a nucleic acid) which are recombined to produce selectable nucleic acids. This format is

5     described in detail in Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed February 5, 1999, USSN 60/118,813 and Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed June 24, 1999, USSN 60/141,049. In brief, selected oligonucleotides are synthesized, ligated and elongated, typically either in a polymerase or ligase-mediated elongation reaction. The

10    technique can be used to recombine homologous or even non-homologous codon-altered nucleic acid sequences.

One advantage of oligonucleotide-mediated recombination is the ability to recombine homologous nucleic acids with low sequence similarity, or even non-homologous nucleic acids. In these low-homology oligonucleotide shuffling methods, one or more set of

15    fragmented nucleic acids (e.g., cleaved codon-modified oligonucleotides, or synthesized codon-modified oligonucleotides) are recombined, e.g., with a with a set of crossover family diversity oligonucleotides. Each of these crossover oligonucleotides have a plurality of sequence diversity domains corresponding to a plurality of sequence diversity domains from homologous or non-homologous nucleic acids with low sequence similarity. The fragmented

20    oligonucleotides, which are derived by comparison to one or more homologous or non-homologous nucleic acids, can hybridize to one or more region of the crossover oligos, facilitating recombination.

When recombining homologous nucleic acids, sets of overlapping family gene shuffling oligonucleotides (which are derived by comparison of homologous nucleic acids

25    that include one or more codon-modified nucleic acid, followed by synthesis of corresponding oligonucleotides) are hybridized and elongated (e.g., by reassembly PCR or ligation), providing a population of recombined nucleic acids, which can be selected for a desired trait or property. The set of overlapping family shuffling gene oligonucleotides includes a plurality of oligonucleotide member types which have consensus region

30    subsequences derived from a plurality of homologous target nucleic acids.

46

Typically, as applied to the present invention, family gene shuffling oligonucleotide which include one or more codon-altered nucleic acid(s) are provided by aligning homologous nucleic acid sequences to select conserved regions of sequence identity and regions of sequence diversity. A plurality of family gene shuffling oligonucleotides are

5      synthesized (serially or in parallel) which correspond to at least one region of sequence diversity.

Sets of fragments, or subsets of fragments used in oligonucleotide shuffling approaches can be provided by cleaving one or more homologous nucleic acids (e.g., with a DNase), or, more commonly, by synthesizing a set of oligonucleotides corresponding to a

10     plurality of regions of at least one nucleic acid (typically oligonucleotides corresponding to a full-length nucleic acid are provided as members of a set of nucleic acid fragments). In the shuffling procedures herein, these cleavage fragments can be used in conjunction with family gene shuffling oligonucleotides, e.g., in one or more recombination reaction to produce recombinant codon-altered nucleic acid(s).

15     Additional oligonucleotide shuffling formats are found in co-filed application by Crameri et al., "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" (Attorney Docket Number 02-296-2US) and in co-filed application by Welch et al., "USE OF CODON VARIED OLIGONUCLEOTIDE SYNTHESIS FOR SYNTHETIC SHUFFLING" (Attorney docket number 02-1007). In particular, these

20     applications provide for tri-nucleotide-based synthesis of degenerate oligonucleotides, thereby providing for codon substitution during oligonucleotide shuffling. In brief, this procedure utilizes tri-nucleotide phosphoramidite chemistry to synthesize oligos, rather than standard mono-nucleotide synthesis. Because codons are altered as a unit, the synthetic scheme of degenerate oligonucleotides is simplified.

25     <u>Additional In Vitro DNA Shuffling Formats</u>

In one embodiment for shuffling DNA sequences *in vitro*, the initial substrates for recombination are a pool of related sequences, *e.g.*, different variant forms, as homologs from different individuals, strains, or species of an organism, or related sequences from the same organism, as allelic variations. The sequences can be DNA or RNA and can be of

30     various lengths depending on the size of the gene or DNA fragment to be recombined or reassembled. Preferably the sequences are from 50 base pairs (bp) to 50 kilobases (kb).

The pool of related substrates are converted into overlapping fragments, e.g., from about 5 bp to 5 kb or more. Often, for example, the size of the fragments is from about 10 bp to 1000 bp, and sometimes the size of the DNA fragments is from about 100 bp to 500 bp. The conversion can be effected by a number of different methods, such as DNase I or

5    RNase digestion, random shearing or partial restriction enzyme digestion, or by oligonucleotide synthesis as in the family oligonucleotide-mediated shuffling methods of crameri et al., discusses *supra*. For discussions of protocols for the isolation, manipulation, enzymatic digestion, and the like, of nucleic acids, *see*, for example, Sambrook *et al.* and Ausubel, both *supra*. The concentration of nucleic acid fragments of a particular length and

10    sequence is often less than 0.1 % or 1% by weight of the total nucleic acid. The number of different specific nucleic acid fragments in the mixture is usually at least about 2, 10, 100, 500 or 1,000 or more.

The mixed population of nucleic acid fragments are converted to at least partially single-stranded form using any of a variety of techniques, including, for example,

15    heating, chemical denaturation, use of DNA binding proteins, and the like (in oligonucleotide mediated methods, this step can be omitted). Conversion can be effected by heating to about 80 °C to 100 °C, more preferably from 90 °C to 96 °C, to form single-stranded nucleic acid fragments and then reannealing. Conversion can also be effected by treatment with a single-stranded DNA binding protein (see Wold (1997) *Annu. Rev. Biochem.* 66:61-92) or *recA*

20    protein (*see, e.g.*, Kiianitsa (1997) *Proc. Natl. Acad. Sci. U S A* 94:7837-7840). Single-stranded nucleic acid fragments having regions of sequence identity with other single-stranded nucleic acid fragments can then be reannealed by cooling to 20 °C to 75 °C, and preferably from 40 °C to 65 °C. Renaturation can be accelerated by the addition of polyethylene glycol (PEG), other volume-excluding reagents or salt. The salt concentration

25    is preferably from 0 mM to 200 mM, more preferably the salt concentration is from 10 mM to 100 mM. The salt may be KCl or NaCl. The concentration of PEG is preferably from 0% to 20%, more preferably from 5% to 10%. The fragments that reanneal can be from different substrates. The annealed nucleic acid fragments are incubated in the presence of a nucleic acid polymerase, such as Taq or Klenow, and dNTP's (*i.e.* dATP, dCTP, dGTP and dTTP). If

30    regions of sequence identity are large, Taq polymerase can be used with an annealing temperature of between 45-65 °C. If the areas of identity are small, Klenow polymerase can

be used with an annealing temperature of between 20-30 °C. The polymerase can be added to the random nucleic acid fragments prior to annealing, simultaneously with annealing or after annealing.

The process of denaturation, renaturation and incubation in the presence of
5   polymerase or ligase of overlapping fragments to generate a collection of polynucleotides containing different permutations of fragments is sometimes referred to as shuffling of the nucleic acid *in vitro*. This cycle is repeated for a desired number of times. Preferably the cycle is repeated from 2 to 100 times, more preferably the sequence is repeated from 10 to 40 times. The resulting nucleic acids are a family of double-stranded polynucleotides of from
10  about 50 bp to about 100 kb, preferably from 500 bp to 50 kb. The population represents variants of the starting substrates showing substantial sequence identity thereto but also diverging at several positions. The population has many more members than the starting substrates. The population of fragments resulting from shuffling is used to transform host cells, optionally after cloning into a vector.

15          In one embodiment utilizing *in vitro* shuffling, subsequences of recombination substrates can be generated by amplifying the full-length sequences under conditions which produce a substantial fraction, typically at least 20 percent or more, of incompletely extended amplification products. Another embodiment uses random primers to prime an entire template DNA to generate less than full length amplification products. The amplification
20  products, including the incompletely extended amplification products are denatured and subjected to at least one additional cycle of reannealing and amplification. This variation, in which at least one cycle of reannealing and amplification provides a substantial fraction of incompletely extended products, is termed "stuttering." In the subsequent amplification round, the partially extended (less than full length) products reanneal to and prime extension
25  on different sequence-related template species. In another embodiment, the conversion of substrates to fragments can be effected by partial PCR amplification of substrates.

In another embodiment, a mixture of fragments is spiked with one or more oligonucleotides. The oligonucleotides can be designed to include precharacterized mutations of a wildtype sequence (e.g., codon modification), or sites of natural variations
30  between individuals or species. The oligonucleotides also typically include sufficient sequence or structural homology flanking such mutations or variations to allow annealing

with the wildtype fragments. Annealing temperatures can be adjusted depending on the length of homology.

In a further embodiment, recombination occurs in at least one cycle by template switching, such as when a DNA fragment derived from one template primes on the homologous position of a related but different template. Template switching can be induced by addition of recA (*see*, Kiianitsa (1997) *supra*), rad51 (*see*, Namsaraev (1997) *Mol. Cell. Biol.* 17:5359-5368), rad55 (*see*, Clever (1997) *EMBO J.* 16:2535-2544), rad57 (*see*, Sung (1997) *Genes Dev.* 11:1111-1121) or other polymerases (*e.g.*, viral polymerases, reverse transcriptase) to the amplification mixture. Template switching can also be increased by increasing the DNA template concentration.

Another embodiment utilizes at least one cycle of amplification, which can be conducted using a collection of overlapping single-stranded DNA fragments of related sequence, and different lengths. Fragments can be prepared using a single stranded DNA phage, such as M13 (*see*, Wang (1997) *Biochemistry* 36:9486-9492). Each fragment can hybridize to and prime polynucleotide chain extension of a second fragment from the collection, thus forming sequence-recombined polynucleotides. In a further variation, ssDNA fragments of variable length can be generated from a single primer by Pfu, Taq, Vent, Deep Vent, UlTma DNA polymerase or other DNA polymerases on a first DNA template (*see*, Cline (1996) *Nucleic Acids Res.* 24:3546-3551). The single stranded DNA fragments are used as primers for a second, Kunkel-type template, consisting of a uracil-containing circular ssDNA. This results in multiple substitutions of the first template into the second. *See*, Levichkin (1995) *Mol. Biology* 29:572-577; Jung (1992) *Gene* 121:17-24.

In some embodiments of the invention, shuffled nucleic acids obtained by use of the recursive recombination methods of the invention, are put into a cell and/or organism for screening. Shuffled genes can be introduced into, for example, bacterial cells, yeast cells, fungal cells vertebrate cells, invertebrate cells or plant cells for initial screening. *Bacillus* species (such as *B. subtilis* and *E. coli* are two examples of suitable bacterial cells into which one can insert and express shuffled genes which provide for convenient shuttling to other cell types (a variety of vectors for shuttling material between these bacterial cells and eukaryotic cells are available; *see*, Sambrook, Ausubel and Berger, *all supra*). The shuffled genes can

50

be introduced into bacterial, fungal or yeast cells either by integration into the chromosomal DNA or as plasmids.

Bacterial, plant, animal and yeast systems are preferred in the present invention. For example, in one embodiment, shuffled genes can be introduced into plant or animal cells for production purposes (it will be appreciated that transgenic plants are, increasingly, an important source of industrial enzymes), or can be introduced into a plant or animal cell for therapeutic purposes. Thus, a transgene of interest can be modified using the recursive sequence recombination methods of the invention *in vitro* and reinserted into the cell for *in vivo/in situ* selection for the new or improved property, in bacteria, eukaryotic cells, or whole eukaryotic organisms.

### In Vivo DNA Shuffling Formats

In some embodiments of the invention, DNA substrate molecules, e.g., those comprising codon modifications relative to a wild-type sequence, are introduced into cells, where the cellular machinery directs their recombination. For example, a library of mutants is constructed and screened or selected for mutants with improved phenotypes by any of the techniques described herein.

The DNA substrate molecules encoding the best candidates are recovered by any of the techniques described herein, then fragmented and used to transfect a plant host and screened or selected for improved function. If further improvement is desired, the DNA substrate molecules are recovered from the host cell, such as by PCR, and the process is repeated until a desired level of improvement is obtained. In some embodiments, the fragments are denatured and reannealed prior to transfection, coated with recombination stimulating proteins such as recA, or co-transfected with a selectable marker such as Neo$^R$ to allow the positive selection for cells receiving recombined versions of the gene of interest. Methods for *in vivo* shuffling are described in, for example, PCT application WO 98/13487 and WO 97/20078. The efficiency of *in vivo* shuffling can be enhanced by increasing the copy number of a gene of interest in the host cells.

### Whole Genome Shuffling

In one embodiment, the selection methods herein are utilized in a "whole genome shuffling" format. An extensive guide to the many forms of whole genome shuffling is found in the pioneering application to the inventors and their co-workers entitled

51

"Evolution of Whole Cells and Organisms by Recursive Sequence Recombination,"
PCT/US99/15972, by del Cardayre et al. Any codon-altered set of nucleic acids can be used
to transform cells, which can then be shuffled by in a whole genome format.

In brief, whole genome shuffling makes no presuppositions at all regarding

5    what nucleic acids may confer a desired property. Instead, entire genomes (e.g., from a
genomic -library, or isolated from an organism) are shuffled in cells and selection protocols
applied to the cells. These genomes can be spiked with any desired set of nucleic acids,
including codon-modified nucleic acids.

### Assays

10   The relevant assay for selection of a desired property of a codon-modified
nucleic acid will depend on the application. Many assays which detect activity for proteins,
receptors, ligands, cells and the like are known. Formats include binding to immobilized
components, cell or organismal viability, production of reporter compositions, and the like.

In the high throughput assays of the invention, it is possible to screen up to

15   several thousand different shuffled variants in a single day. In particular, each well of a
microtiter plate can be used to run a separate assay, or, if concentration or incubation time
effects are to be observed, every 5-10 wells can test a single variant. Thus, a single standard
microtiter plate can assay about 100 (e.g., 96) reactions. If 1536 well plates are used, then a
single plate can easily assay from about 100- about 1500 different reactions. It is possible to

20   assay several different plates per day; assay screens for up to about 6,000-20,000 different
assays (i.e., involving different nucleic acids, encoded proteins, concentrations, etc.) is
possible using the integrated systems of the invention. More recently, microfluidic
approaches to reagent manipulation have been developed, e.g., by Caliper Technologies
(Mountain View, CA).

25   In one aspect, library members, e.g., cells, viral plaques, spores or the like, are
separated on solid media to produce individual colonies (or plaques). Using an automated
colony picker (e.g., the Q-bot, Genetix, U.K.), colonies or plaques are identified, picked, and
up to 10,000 different mutants inoculated into 96 well microtitre dishes containing two 3 mm
glass balls/well. The Q-bot does not pick an entire colony but rather inserts a pin through the

30   center of the colony and exits with a small sampling of cells, (or mycelia) and spores (or
viruses in plaque applications). The time the pin is in the colony, the number of dips to

inoculate the culture medium, and the time the pin is in that medium each effect inoculum
size, and each can be controlled and optimized. The uniform process of the Q-bot decreases
human handling error and increases the rate of establishing cultures (roughly 10,000/4 hours).
These cultures are then shaken in a temperature and humidity controlled incubator. The

5    glass balls in the microtiter plates act to promote uniform aeration of cells and the dispersal
of mycelial fragments similar to the blades of a fermenter. Clones from cultures of interest
can be cloned by limiting dilution. As also described supra, plaques or cells constituting
libraries can also be screened directly for production of proteins, either by detecting
hybridization, protein activity, protein binding to antibodies, or the like.

10            The ability to detect a subtle increase in the performance of a shuffled library
member over that of a parent strain relies on the sensitivity of the assay. The chance of
finding the organisms having an improvement is increased by the number of individual
mutants that can be screened by the assay. To increase the chances of identifying a pool of
sufficient size, a prescreen that increases the number of mutants processed by, e.g., 10-fold

15    can be used. The goal of the primary screen is to quickly identify mutants having equal or
better product titres than the parent strain(s) and to move only these mutants forward to liquid
cell culture for subsequent analysis.

              A number of well known robotic systems have also been developed for
solution phase chemistries useful in assay systems. These systems include automated

20    workstations like the automated synthesis apparatus developed by Takeda Chemical
Industries, LTD. (Osaka, Japan) and many robotic systems utilizing robotic arms (Zymate II,
Zymark Corporation, Hopkinton, Mass.; Orca, Hewlett-Packard, Palo Alto, Calif.) which
mimic the manual synthetic operations performed by a scientist. Any of the above devices
are suitable for use with the present invention, e.g., for high-throughput screening of

25    molecules encoded by codon-altered nucleic acids. The nature and implementation of
modifications to these devices (if any) so that they can operate as discussed herein with
reference to the integrated system will be apparent to persons skilled in the relevant art.

              High throughput screening systems are commercially available (see, e.g.,
Zymark Corp., Hopkinton, MA; Air Technical Industries, Mentor, OH; Beckman

30    Instruments, Inc. Fullerton, CA; Precision Systems, Inc., Natick, MA, etc.). These systems
typically automate entire procedures including all sample and reagent pipetting, liquid

dispensing, timed incubations, and final readings of the microplate in detector(s) appropriate for the assay. These configurable systems provide high throughput and rapid start up as well as a high degree of flexibility and customization.

The manufacturers of such systems provide detailed protocols the various high

5    throughput. Thus, for example, Zymark Corp. provides technical bulletins describing screening systems for detecting the modulation of gene transcription, ligand binding, and the like. Microfluidic approaches to reagent manipulation have also been developed, e.g., by Caliper Technologies (Mountain View, CA).

Optical images viewed (and, optionally, recorded) by a camera or other

10   recording device (e.g., a photodiode and data storage device) are optionally further processed in any of the embodiments herein, e.g., by digitizing the image and/or storing and analyzing the image on a computer. A variety of commercially available peripheral equipment and software is available for digitizing, storing and analyzing a digitized video or digitized optical image, e.g., using PC (Intel x86 or pentium chip- compatible DOS™, OS2™

15   WINDOWS™, WINDOWS NT™ or WINDOWS95™ based machines), MACINTOSH™, or UNIX based (e.g., SUN™ work station) computers.

One conventional system carries light from the assay device to a cooled charge-coupled device (CCD) camera, in common use in the art. A CCD camera includes an array of picture elements (pixels). The light from the specimen is imaged on the CCD.

20   Particular pixels corresponding to regions of the specimen (e.g., individual hybridization sites on an array of biological polymers) are sampled to obtain light intensity readings for each position. Multiple pixels are processed in parallel to increase speed. The apparatus and methods of the invention are easily used for viewing any sample, e.g., by fluorescent or dark field microscopic techniques.

25   Software elements for manipulating strings of characters which correspond to codon-modified nucleic acids can be used to direct synthesis of oligonucleotides relevant to shuffling of codon-modified nucleic acids. Integrated systems comprising these and other useful features, e.g., a digital computer with additional features such as high-throughput liquid control software, image analysis software, data interpretation software, a robotic liquid

30   control armature for transferring solutions from a source to a destination operably linked to the digital computer, an input device (e.g., a computer keyboard) for entering data to the

54

digital computer to control high throughput liquid transfer by the robotic liquid control armature an image scanner for digitizing label signals from labeled assay components, or the like are a feature of the invention.

In one aspect, the invention provides an integrated system comprising a

5    computer or computer readable medium comprising a database having at least two artificial homologous codon-altered nucleic acid sequence strings, and a user interface allowing a user to selectively view one or more sequence strings in the database. As discussed theroughout, there are a variety of sequence database programs for aligning and manipulating sequences. In addition, standard text manipulation software such as word processing software (e.g.,

10   Microsft Word™ or Corel Wodperfect™) and database software (e.g., spreadsheet software such as Microsoft Excel™, Corel Quattro Pro™, or database programs such as Microsoft Access™ or Paradox™) can be used in conjuction with a user interface (e.g., a GUI in a standard operating system such as a Windows, Macintosh or LINUX system) to manipulate strings of characters. Specialized alignment programs such as BLAST can also be

15   incorporated into the systems of the invention for alignment of codon-altered nucleic acids (or corresponding character strings).

In addition to the integrated system elements mentioned above, the integrated system can also include an automated oligonucleotide synthesizer operably linked to the computer or computer readable medium. Typically, the synthesizer is programmed to

20   synthesize one or more oligonucleotide comprising one or more subsequence of one or more of the at least two artificial homologous codon-altered nucleic acids.

Modifications can be made to the method and materials as hereinbefore described without departing from the spirit or scope of the invention as claimed, and the invention can be put to a number of different uses, including:

25   The use of an integrated system to test shuffled codon-modified DNAs, including in an iterative process.

An assay, kit or system utilizing a use of any one of the selection strategies, materials, components, methods or substrates hereinbefore described. Kits will optionally additionally comprise instructions for performing methods or assays, packaging materials,

30   one or more containers which contain assay, device or system components, or the like.

In an additional aspect, the present invention provides kits embodying the methods and apparatus herein. Kits of the invention optionally comprise one or more of the following: (1) a shuffled codon-modified component as described herein; (2) instructions for practicing the methods described herein, and/or for operating the selection procedure herein;

5  (3) one or more assay component; (4) a container for holding nucleic acids or enzymes, other nucleic acids, transgneic plants, animals, cells, or the like, (5) packaging materials and (6) software fixerd in a computer readable medium comprising sequences corresponding to one or more codon-altered nucleic acid character string.

In a further aspect, the present invention provides for the use of any

10  component or kit herein, for the practice of any method or assay herein, and/or for the use of any apparatus or kit to practice any assay or method herein.

While the foregoing invention has been described in some detail for purposes of clarity and understanding, it will be clear to one skilled in the art from a reading of this disclosure that various changes in form and detail can be made without departing from the

15  true scope of the invention. For example, all the techniques and materials described above can be used in various combinations. All publications and patent documents cited in this application are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication or patent document were so individually denoted.

56

WHAT IS CLAIMED IS:

      1.    A method of making codon altered nucleic acids, the method comprising:
(i) providing a first nucleic acid sequence, which nucleic acid sequence encodes a first
5   polypeptide sequence;
(ii) providing a plurality of codon altered nucleic acid sequences, each of which encode the
first polypeptide or a modified form thereof; and,
(iii) recombining the plurality of codon-altered nucleic acid sequences to produce a target
codon altered nucleic acid, which target codon altered nucleic acid encodes a second protein.

10      2.    The method of claim 1, wherein at least one of the plurality of codon
altered nucleic acid sequences does not hybridize to the first nucleic acid under stringent
hybridization conditions.

      3.    The method of claim 1, further comprising shuffling a nucleic acid
comprising a subsequence consisting of the first nucleic acid, or a substantially identical
15   variant thereof, with one or more of the plurality of codon altered nucleic acids, or with the
target codon altered nucleic acid.

      4.    The method of claim 1, the method further comprising the step of:
(iv) screening the second protein for a structural or functional property.

      5.    The method of claim 1, the method further comprising the steps of:
20   (iv) screening the second protein for a structural or functional property, and,
(v) comparing the structural or functional property of the second protein to a structural or
functional property of the first protein.

      6.    The method of claim 1, wherein the second polypeptide has a structural
or functional property equivalent or superior to the first polypeptide.

25      7.    The method of claim 1, wherein the first and second polypeptide are
homologous.

8.    The method of claim 1, wherein the plurality of codon altered nucleic acids comprise a library of codon altered nucleic acids.

9.    The method of claim 1, wherein the plurality of codon altered nucleic acids comprise a library of codon altered conservatively modified nucleic acids.

10.    A library of codon altered conservatively modified nucleic acids produced by the method of claim 9.

11.    The method of claim 1, wherein the plurality of codon altered nucleic acids comprise a library of codon altered non-conservatively modified nucleic acids.

12.    A library of codon altered conservatively modified nucleic acids produced by the method of claim 11.

13.    The method of claim 1, wherein the plurality of codon altered nucleic acids is derived from a plurality of forms of the first nucleic acid.

14.    The method of claim 1, wherein the plurality of codon altered nucleic acid sequences comprise at least three codon altered nucleic acids.

15.    The method of claim 1, wherein the plurality of codon altered nucleic acid sequences comprise one or more of the following structural features:

(a) codon usage divergence for each of the codon altered nucleic acids of 50% or more as compared to the first nucleic acid;

(b) codon usage divergence for each of the codon altered nucleic acids of 75% or more as compared to the first nucleic acid;

(c) codon usage divergence for each of the codon altered nucleic acids of 90% or more as compared to the first nucleic acid;

(d) maximal codon usage divergence for each of the codon altered nucleic acids as compared to the first nucleic acid;

(e) non-overlapping non-conservative substitutions in each of the codon altered nucleic acids as compared to the first nucleic acid;

(f) a lack of high stringency hybridization between one or more of the codon altered nucleic acid and the first nucleic acid; and,

(g) modification of the codons of one or more of the codon altered nucleic acids to provide one or more different hydrophobic core residue for an encoded polypeptide as compared to the first polypeptide.

16. The method of claim 1, wherein the percent identity between the second protein and the first protein is lower than the percent identity between two of the plurality of codon altered nucleic acids.

17. The method of claim 1, wherein the first nucleic acid encodes a protein selected from: EPO, G-CSF, a viral envelope protein, a cytokine, and a phosphatase.

18. The method of claim 1, wherein the first nucleic acid sequence or the codon altered nucleic acid sequences are isolated nucleic acids.

19. The target codon altered nucleic acid produced by the method of claim 1.

20. The method of claim 1, wherein the first nucleic acid sequence or the codon altered nucleic acid sequences are nucleic acids present in cells.

21. The cells produced by the method of claim 20.

22. The method of claim 1, wherein each of the codon altered nucleic acid sequences comprises at least two nucleotide differences when compared to the first nucleic acid.

23. The method of claim 1, further comprising introducing the target codon altered nucleic acid into a cell, or into a vector or virus.

24. The cell, vector or virus produced by the method of claim 23.

25. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a viral genome to produce an attenuated virus.

26. The attenuated virus produced by the method of claim 25.

27. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a viral genome to produce an attenuated virus, which attenuated virus produces an immune response upon infection by the virus in a mammal.

28. The attenuated virus produced by the method of claim 27.

29. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a retroviral genome to produce an attenuated retrovirus, which attenuated retrovirus produces an immune response upon infection by the retrovirus in a mammal.

30. The attenuated retrovirus produced by the method of claim 29.

31. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a viral genome to produce an viral vector.

32. The viral vector produced by the method of claim 31.

33. The method of claim 1, wherein the target codon altered nucleic acid is recombined with a portion of a viral genome to produce an viral vector, which vector requires *trans* complementation for replication, and which vector has a reduced rate of reversion to a replicative form as compared to a corresponding viral vector which lacks a subsequence corresponding to the target codon altered nucleic acid.

34. The viral vector produced by the method of claim 33.

35. The method of claim 33, wherein the vector comprises viral elements from one or more of: a lentivirus, an adenovirus, a herpes virus, and an adeno-associated virus.

36. A method of making a library of codon-altered nucleic acids, the method comprising:

(i) selecting a first nucleic acid sequence, which nucleic acid sequence encodes a first polypeptide sequence; and,

(ii) making a plurality of codon altered nucleic acid sequences, each of which encode the first polypeptide or a modified form thereof, wherein the plurality of codon altered nucleic acids comprise the library.

37. A codon-altered library made by the method of claim 36.

38. The library of claim 37, wherein said library comprises at least 2 codon altered nucleic acids.

39. The library of claim 37, wherein said library comprises at least 5 codon altered nucleic acids.

40. The library of claim 37, wherein said library comprises at least 10 codon altered nucleic acids.

41. The library of claim 37, wherein said library comprises at least 100 codon altered nucleic acids.

42. A kit comprising the library of claim 25 and one or more of: a container and instructional materials providing method step instructions for recombining two or more of members of the library.

43. The method of claim 41, further comprising recombining said plurality of codon altered nucleic acids to produce a shuffled codon-altered library.

44. The codon altered library made by the method of claim 43.

45. The method of claim 36, wherein said nucleic acids encode a protein selected from EPO, a cytokine, a phosphatase, and a viral envelope protein.

46. A composition comprising a plurality of codon altered nucleic acids, each of which encode a first polypeptide or a modified form thereof.

47. A library of codon-altered nucleic acids, comprising a plurality of codon-altered nucleic acids derived from a plurality of homologous nucleic acids.

48.  The library of claim 47, wherein said plurality of codon altered nucleic acids recombine in vitro at an increased rate compared to said plurality of homologous nucleic acids.

49.  The library of claim 47, wherein the level of identity among said plurality of codon-altered nucleic acids is at least as high as among a plurality of polypeptides encoded by said plurality of homologous nucleic acids.

50.  An integrated system comprising a computer or computer readable medium comprising a database having at least two artificial homologous codon-altered nucleic acid sequence strings, and a user interface allowing a user to selectively view one or more sequence strings in the database.

51.  The integrated system of claim 50, further comprising an automated oligonucleotide synthesizer operably linked to the computer or computer readable medium, which synthesizer is programmed to synthesize one or more oligonucleotide comprising one or more subsequence of one or more of the at least two artificial homologous codon-altered nucleic acids.

EPO cDNA (Monkey)

FIG.5A

Translation of monkey EPO cDNA

Sau3A
GATCCCCCCCCCCCCCGGATACCCCCTCCCCGCCAGCCCTGCGAGCTCGCCCCCGCCT
CCCTCAATTCCCGGGGATGACCACTCCCCGGTGGGTCACCCGCACCCGGGGCTAGCTCCCTGAG

                -22  Gly Val His Glu Cys Pro Ala Trp
                Met Gly Val His Glu Cys Pro Ala Trp
GGGCCCCGGCCCCAGGCGGGCAGATC GGG GTG CAC GAA TGT CCT GCC TGG

      -3  -1                    -10  Ser Leu Leu Ser Leu Leu Val Ser Leu Pro Leu Gly Leu Pro
Leu Trp Leu Leu Ser Leu Leu Val Ser Leu Pro Leu Gly Leu Pro
CTG TGG CTT CTC CTG CTC GTC TCC CTC CCT CTG GGC CTC CCA

      -1  1              10  Val Pro Gly Ala Pro Pro Arg Leu Ile Cys Asp Ser Arg Val Leu
Val Pro Gly Ala Pro Pro Arg Leu Ile Cys Asp Ser Arg Val Leu
GTC CCG GGC GCC CCA CCA CGC CTC ATC TGT GAC AGC CGA GTC CTG

                20  Glu Arg Tyr Leu Leu Glu Ala Lys Glu Ala Glu Asn Val Thr Met
Glu Arg Tyr Leu Leu Glu Ala Lys Glu Ala Glu Asn Val Thr Met
GAG AGG TAC CTC TTG GAG GCC AAG GAG GCC GAG AAT GTC ACG ATG

              30                      40  Gly Cys Ser Glu Ser Leu Asn Glu Asn Ile Thr Val Pro
Gly Cys Ser Glu Ser Leu Asn Glu Asn Ile Thr Val Pro
GGC TGT TCC GAA AGC CTG AAC GAG AAT ATC ACC GTC CCA

Fig

EPO wobble gene

```
                10          20          30          40          50
        GCCCCACCACGCCTCATCTGTCACAGCCGAGTCCTGGAGAGGTACCTCTT   EPO coding
        ..G...T...TA..GT.G...A...C...TTCGA..G...GT.A...AC..C...TT..GC.   EPO wobble
                                        INTRON

                60          70          80          90         100
    51  GGAGGCCAAGGAGGCCGAGAATATCACC...C.C...G....C..A...C...T...TT   EPO coding
    51  C..A...G...A...A...G...A...C...   EPO wobble

               110         120         130         140         150
   101  GCTTCAATGAGAATATCACTCTCCCAGACACCAAAGTTAATTTCTATGCC   EPO coding
   101  CGGC..C...C...A...   EPO wobble
                        INTRON           INTRON/INTRON
                                                                200
        TGGAGGAGCATGGAGTCTGGGCAGCAGGCCCTAGAAGTCTGGCAGGGCCT  EPO coding
        ...AC..   ..A...   A..GT.   EPO wobble

               310         220         230         240         250
   201  CGGCCTGCTGTCCGAAGCGTGTCCTGCGGGGCAGGCCCTGTTGGTCAACT   EPO coding
   201  A..GT.AT.AAGC...G...A...GT.AA.A...G..A..GT.AC.C...G...TA   EPO wobble

               260         270         280         290         300
   251  CTTCCCAGCCCTGGGAGCCCTGCAGCTCATGTGGATAAAGCCGTCAGT   EPO coding
   251  GCAGT...A...G...GT.A..AT.A...C...C...G...GTCA   EPO wobble
                                                    INTRON

               110         320         330         340         350
   301  GGCCTTCGCAGCCTCACCACTCTGCTTCCGGCTGTGGAGCCTAGAAGGA   EPO coding
   301  ..GT.AA..GTCGT.G..G..AT.AT.AA.A..AT.A..T..G..A..A...   EPO wobble

               360         370         380         390         400
   351  AGCCATCTCCCCTCCAGATGCGGCCTCAGCTGCTCCACTCCGAACAATCA   EPO coding
   351  G..G..AAGT..A..T..C..C..GAGT..A..A..TT.GA.G..T..A.   EPO wobble

               410         420         430         440         450
   401  CTGCTGACACTTTCCGCAAACTCTTCCGAGTCTACTCCAATTTCCTCCGG   EPO coding
   401  .A..A..T..A..TA.G..GT.G..TA.G..G..TAGT..C...TT.GA.A   EPO wobble
                        INTRON/INTRON
                                                               500
        GGAAAGCTGAAGCTGTACACAGGGGAGCCCTGCAGGACAGGGGACAGATG  EPO coding
        ..T..ATG.A..AT.A..T..T...C..A...G..TC.C...T..C..TC.T.A   EPO wobble

                                                    EPO coding
                                                    EPO wobble

   501  A                                            EPO coding
   501  G                                            EPO wobble
```

Fig 2

Alignment Report of epo homologs.MEG, using Clustal method with PAM250 residue weight table.    Page 1

Thursday, August 06, 1998 5:43 PM

```
          A P P R L I C D S R V L E R Y L L E A K E A E N V T M G C A E G C S F S E N I T V P D T K V N F Y A   Majority

                10            20            30            40            50

      A P P R L I C D S R V L E R Y L L S A K E N E N I T T G C A E H C S L N E N I T V P D T K V N F Y A   human EPO
 1    A P P R L I C D S R V L E R Y L L S A K E N E N V T M . . . . G C . . . . . . . . . . . . . . . . . .   monkey epo
38    . . . . . . . . V . . . . . . L . G . R . . . V . M . . . . G . . F S . . . . . . . . . . . . . . T   cat epo
37    . . . . . . . . . . . . . . . L . . . . . . V . M . . . G P R . S . . . . . . . . . . . . . . . . . .   rat epo
27    . . . . . . . . . . . . . . . L . . . . . . V . M . . . G P R . S . . . . . . . . . . . . . . . . . .   cow epo
26    . . A . . . . . . . . . . . . . . . R . . . A . M . . . . G . . . . . . . . . . . . . . . . . . . T   dog epo
33    . . A . . . . . . . . . . . . . V . M . . . . G . . S . . . . . . . . . . . . . . . . . . . . . . . .   pig epo
23    . . . . . . . . . . . . . . . . . R . . . . . . . . . . G . . F S . . . . . . . . . . . . . . . . . .   sheep epo
18    . . . . . . . . . . . . . . . L . . . . R . . V . M . . . G P R . S . . . . . . . . . . . . . . . . . .   mouse epo
27
```
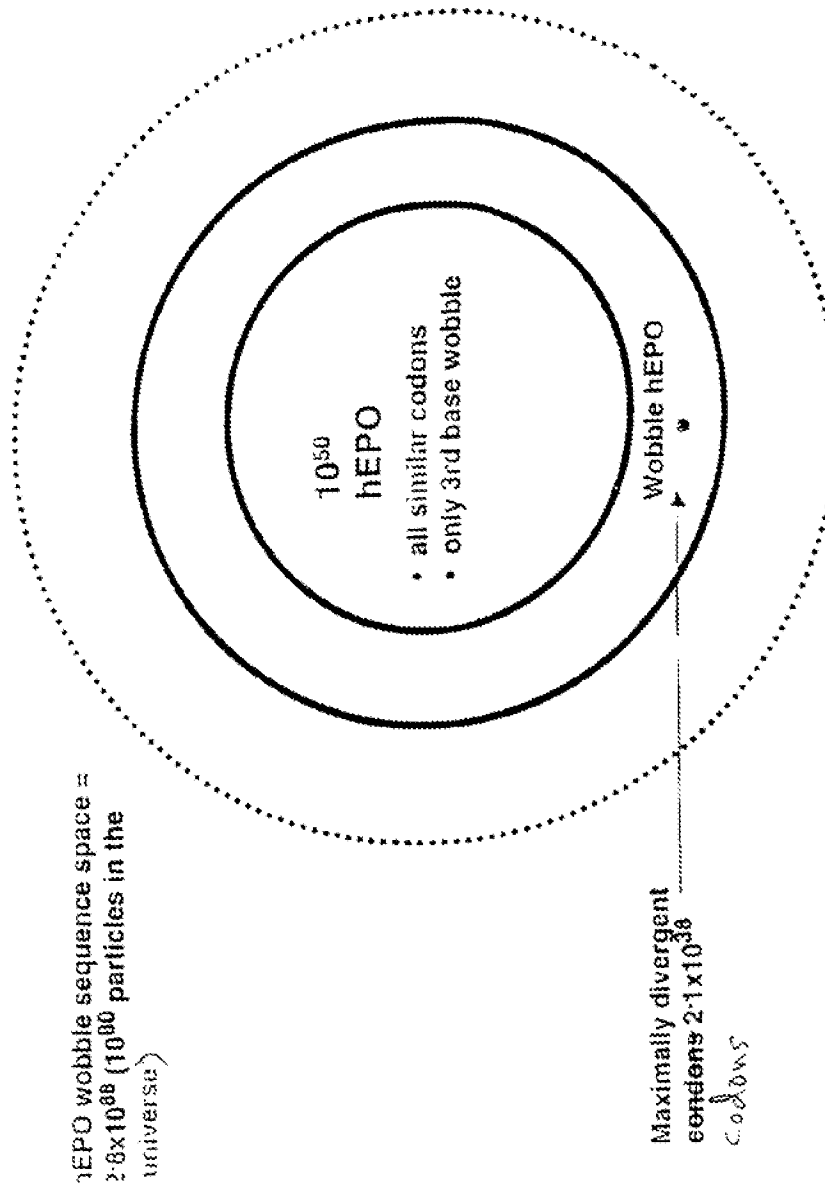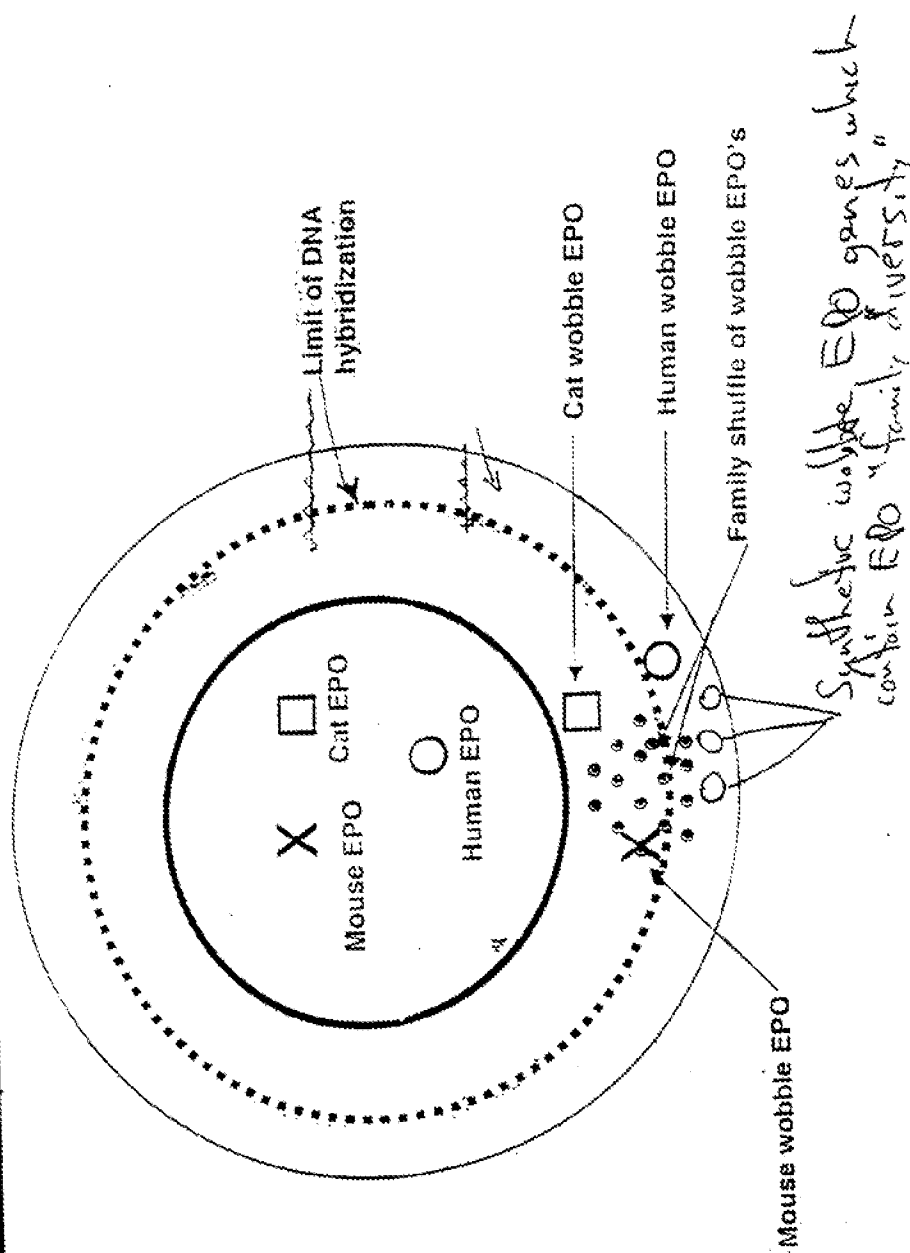
```
          W K R M E V G Q Q A V E V W Q G L A L L S E A I L R G Q A L L A N S S Q P W E A L Q L H V D K A V S   Majority

                60            70            80            90            100

      W K R M E V G Q Q A V E V W Q G L A L L S E A V L R G Q A L L V N S S Q P W E P L Q L H V D K A V S   human EPO
151   W K R M E V G Q Q A V E V W Q G L A L L S E A V L R G Q A V V A . . . . . . . F . . . . M . . . X .   monkey epo
73    . . . . . . . . . . . . . . . . . . . . . . . X . . . . A . . . . S . T . . . . . . . . . . X .   cat epo
37    . . . . . D . . . . . . . . . . . . . . . . . . . . . . . . . . S . . . . . . . . X . . . . . X .   rat epo
77    . . . K . E S . . . . . . . . . . . . . . . . L . Q A . . Q A . . . . C . A . R . . . . . . . . . .   cow epo
76    . . . . . Q . . . . . . . . . . . . . . . . . X . . . . A . A . . S . . P . . . . . . . . . . . . .   dog epo
23    . . . D . . . . . . . . . . . . . . . . . . . X . . . . A . A . . S . . A . . . . . . . . . . . . .   pig epo
73    . . . . . . . . . . . . . . . . . . . . . . . I F . . . A . A . . C . A . R . . . . . . . . . . . .   sheep epo
78    . . . Q . . . . . . . . . . . . . . . . S . . X . Q A . . A . . . . P . T . . . . . X . . . . X .   mouse epo
77
```

```
          G L R S L T T L L R A L G A Q K E A I S P P D A T P - - A A P L R T F T V D T L C K L F R V Y S N F   Majority

                110           120           130           140           150

      G L R S L T T L L R A L G A Q K E A I S P P D A A S - - A A P L R T I T A D T F R K L F R V Y S N F   human EPO
301   G L R S L T T L L R A L G A Q K E A I S P P D A A S - - A A P L R T . T . . . . C . . . . . Y . . .   monkey epo
128   T . . . . T . . . . . . . . . . . . . . . T . L . E . T - S . . . . . . . F . V . . L C . . . . . X . . .   cat epo
127   S . . . . . S . . . . . . . Y . . . . . L K . . . . T Q . . . . . . L . . . . . G . . . . . . . . . .   rat epo
127   . . . . . . S . . . . . . . . . . . . . . . . . T P S A . . . . . A F . V . A L S . . . . I . . . .   cow epo
126   . . . . . . S . . . . . . . . . . . . . . . . T P S A . . . . C . C . . . . . C G . . . . I . . . .   dog epo
133   S . . . . . . . . . . . . . . . . . . . . . . . S F . . . . . C . . . . . . . C G . . . . N . . .   pig epo
123   . . . . . . S . . . . . . . . . . . . . . . . T P S A . . . . I F . V . A L S . . . . I . . . .   sheep epo
128   . . . . . . S . . . . . . . . . . . . . L H . . . T T P - - P . . . L . V . . . C . . . . . A . . .   mouse epo
127   . . . . . . S . . . . . . . . Y . . . . . L H . . . T T P - - P . . . L . V . . . C . . . . . A . . .
                                                                                            Majority
```

```
          L R G K L T L Y T G E A C R R G D R   Majority

                160

      L R G K L K L Y T G E A C R T G D R   human EPO
445   L R G K L K L Y T G E A C R . G D R . . . R Y   monkey epo
175   . . . . . . . . . . . . . . . . . R Y   cat epo
175   . . . . . T . . . . . . . . . . . R .   rat epo
175   . . . . . . . . . . . . . . . . . R .   cow epo
175   . . . . . . T . . . . . . . . . . R .   dog epo
171   . . . . . . . . . . . . . . . . . . .   pig epo
173   . . . . . . T . . . . . . . R R . . .   sheep epo
177   . . . . . . . . . . . . . . . . . . V . R .   mouse epo
175
```

Decoration 'Decoration #1': Hide (as '.') residues that match human EPO exactly.

*52/168 residues can be changed by this family shuffling*

*region of relative conservation*

Fig. 3

# Human EPO Wobble Sequence Space

hEPO wobble sequence space =
2·8×10⁸⁸ (10⁸⁰ particles in the
universe)

$10^{50}$
hEPO
• all similar codons
• only 3rd base wobble

Wobble hEPO

Maximally divergent
codons 2·1×10³⁸
codons

Fig. 4

Fig. 5

# Mammalian EPO Family-Wobble Sequence Space



Limit of DNA hybridization

Cat wobble EPO

Human wobble EPO

Family shuffle of wobble EPO's

Cat EPO

Human EPO

Mouse EPO

Mouse wobble EPO

Synthetic wobble EPO genes which contain EPO "family diversity"

hment Report of gcsf homologs.MEG, using Clustal method with PAM250 residue weight table. *Page 1*

rsday, August 06, 1998 9:06 PM

Fig. 6

# G-CSF Family Diversity

TPLGPASSLPQSFLLKCLEQVRKIQGDGAALQEKLCAT
A    STRP  R    S     M   V  ASNTE L R    A
V    P  SG              R   SV   T      Q
I    L                      S

YKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCL
HQ   Q      M FR A    L   QPS   S    QTS
                           K         RQ
                                     MK

SQLHSQLFLY QGLLQALEGISPELGPTLQTLQLDVAD
D    Q   V         G         A   M H  TN
N    C          S           A  A     I  TD
H                                 I   L

FATTIWQQMEELGMAPALQPTQGAMPAFASAFQRRAG
E   TN  L    DVRV TVP    SFV T T
            N  A     SL       I
            S

QVLVASHLQSFLEVSYRVLRHLAQ
T  AVQ HR  GLARLG HYFTE
I   N G   T NHA   PK
  T Y             H

Fig. 7

Fig. 8

G-CSF

~~G-CSG~~ Family Diversity



```
  TPLGPASS POS   KC EOVRK OGDGAA QEK CAT
A     S  T P  R         M   V A NTE  L  R   A
V     P  S G              K   SV       T
I        L                              Q

YK CHPEE   GHS G PWA SSCPSQA DLAGC
HQ         Q   M  FR A  L  QPS   G S   S    QTS
                        K                   RQ
                                            MK

SQ HSG FNY OG  QA EG SPE GPT D Q D AD
D     U   V     C   A      S AA A    MH    TN
N                                     I    TD
R                                     L

ATTWQQ EE G APA PTQGA P FASAFQRRAG
L IN  L     DVRV    TVP    STV  T  T
            N    A    SL        I
            S

QV YASH QC ER SYRV RH AQ
T  AV  Q HR  GLARLG TFTE
   I   N  G   T HHA      PK
   T   Y                 R
```
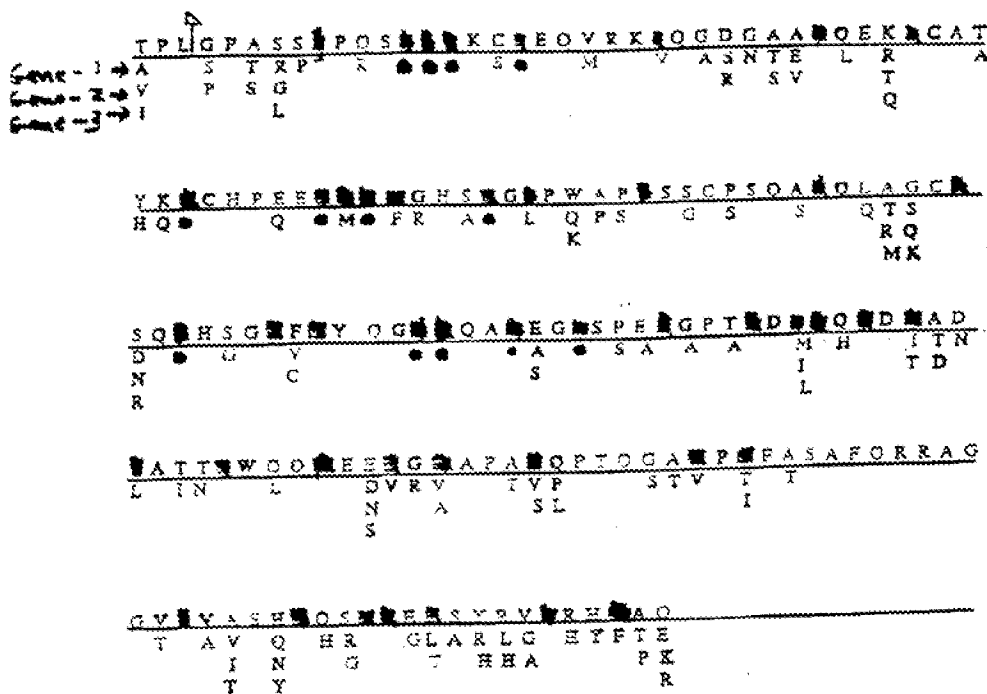
• = hydrophobic core residues in conserved stretches of the protein



wild d.or     Maximally     Non-hybridizing

repacked hydrophobic core
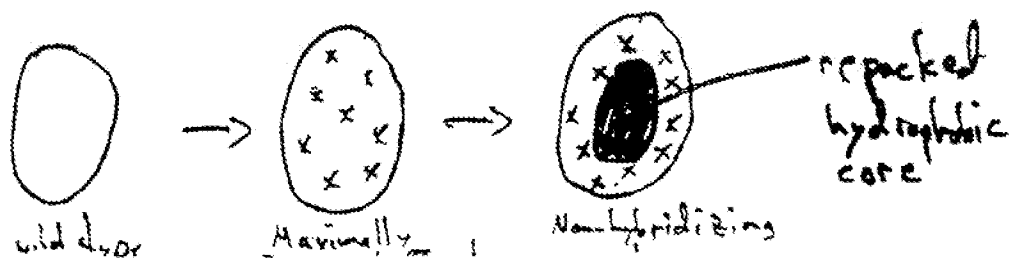
Fig. 9

Strategy for G-CSF Evolution

Fig. 10

Assembling of Alkaline Phosphatase

Fig. 11

## Vaccination with evolution-defective viruses

### Preventing viral immune escape by altering the mutation spectrum through massive codon usage alteration

HIV1

gag          pol          env

HIV1-mutant

gag          pol          env
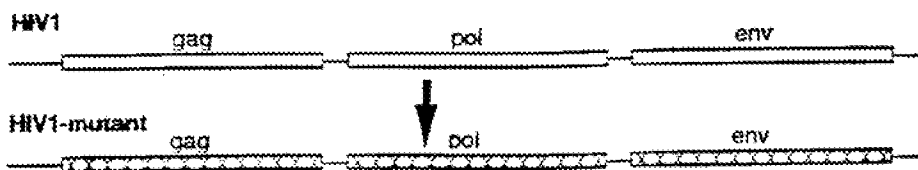
Same current protein sequence,
but mutations result in a different future protein sequence,
a different and unsophisticated mutant cloud.

| Most Effective | | Least Effective | |
|---|---|---|---|
| Ser : TCX ⟷ AGY | | Phe, Tyr, Cys, His, Asn, Asp: | |
| Arg : CGX ⟷ AGR | | - - C ⟷ - - T | |
| Leu : CTX ⟷ TTR | | Lys, Gln, Glu | |
| | | - - A ⟷ - - G | |
| Moderately Effective | | | |
| Thr, Pro, Val, Ala, Gly: | | Unchangeable: | |
| - - R ⟷ - - Y | | Trp : TGG | |
| Ile : ATY ⟷ ATA | | Met : ATG | |

Fig.  12

```
Ser  TCX  ─AGY
Arg  CGX  ─AGR
Leu  CTX  ─TTR
```

Single mutant spectrum

```
        ┌─ CCA  Pro        Arg  CGT ─┐
        ├─ GCA  Ala        Gly  GGT ─┤
        ├─ ACA  Thr        Cys  TGT ─┤
        ├─ TGA  stop       Thr  ACT ─┤
TCA  ───┤─ TTA  Leu        Ile  ATT ─┤─── AGT
(Ser)   ├─ TAA  stop       Asn  AAT ─┤    (Ser)
        ├─ TCT  Ser        Ser  AGC ─┤
        ├─ TCG  Ser        Arg  AGG ─┤
        └─ TCC  Ser        Arg  AGA ─┘
```

Common: 11% Ser, 11% Thr

78% of mutations result in different amino acids

Fig. 13

Figure 14

# Vaccination with evolution-defective viruses

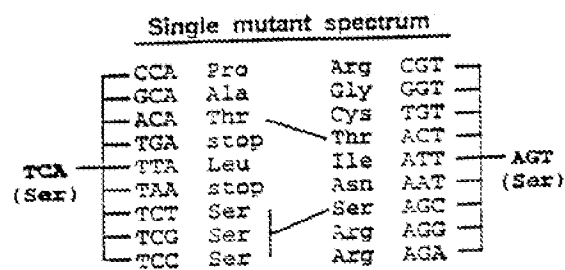## Preventing viral immune escape by altering the mutation spectrum through massive codon usage alteration

HIV1          gag              pol              env

HIV1-mutant   gag              pol              env

Same current protein sequence,
but mutations result in a different future protein sequence,
a different and unsophisticated mutant cloud.

Maximizing the difference requires the following changes:

| Most Effective | Least Effective |
|---|---|
| Ser : TCX ⟷ AGY<br>Arg : CGX ⟷ AGR<br>Leu : CTX ⟷ TTR | Phe, Tyr, Cys, His, Asn, Asp:<br><br>· · C ⟷ · · T<br><br>Lys, Gln, Glu<br><br>· · A ⟷ · · G |

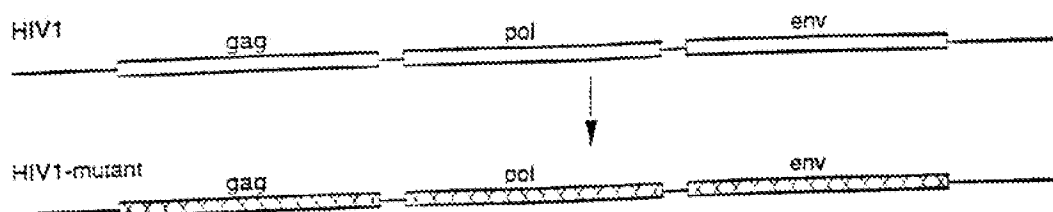| Moderately Effective | |
|---|---|
| Thr, Pro, Val, Ala, Gly:<br><br>· · R ⟷ · · Y<br><br>Ile : ATY ⟷ ATA | Unchangeable:<br><br>Trp : TGG<br>Met : ATG |

Fig. 15

Vaccination with evolution-defective viruses

Preventing viral immune escape by altering the mutation spectrum
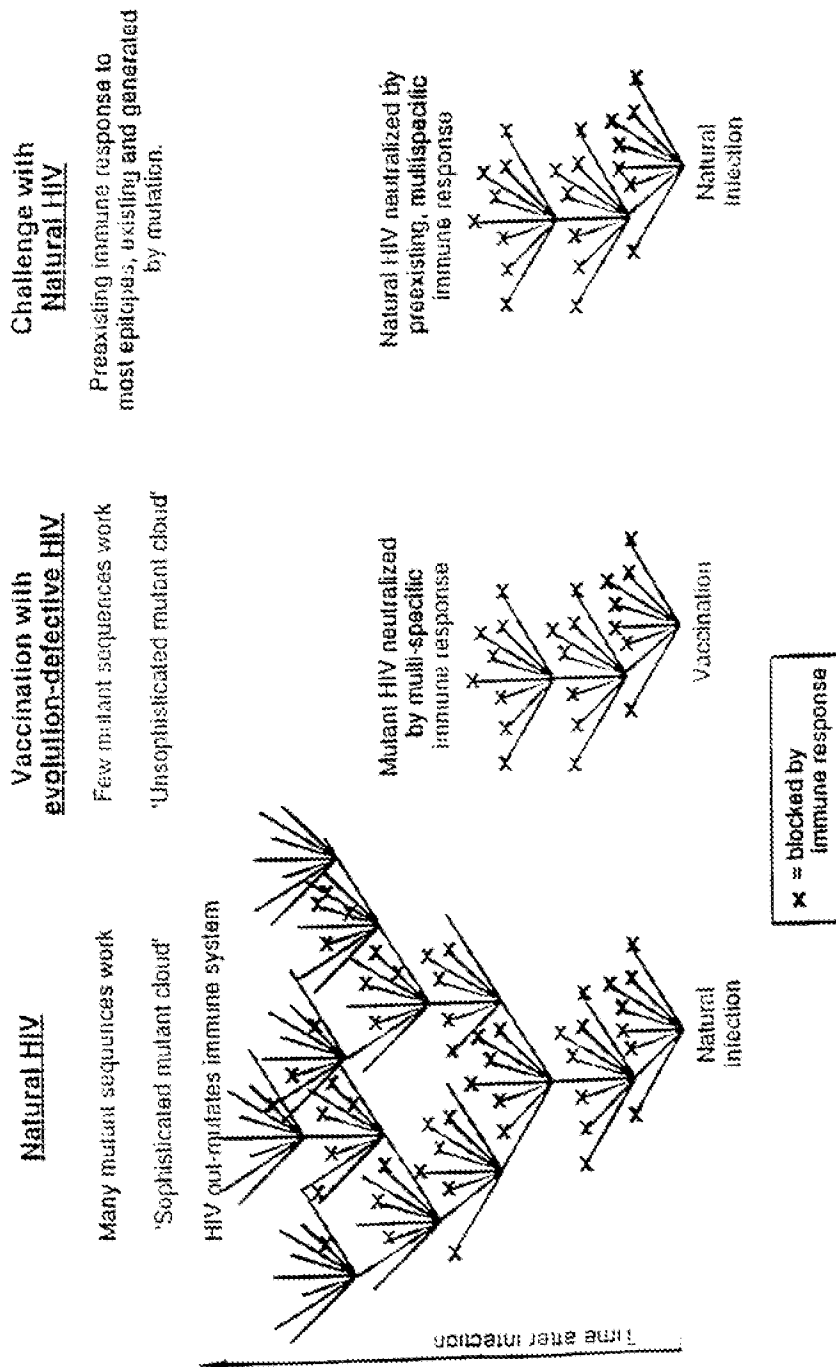through massive codon usage alteration

Fig. 16 A

## Arg

| AGA | | CGT | | AGA | | CGC | | AGA | | CGA | | AGA | | CGG | |
|-----|---|-----|---|-----|---|-----|---|-----|---|-----|---|-----|---|-----|---|
| TGA | . | AGT | S | TGA | . | AGC | S | TGA | . | AGA | S | TGA | . | AGG | S |
| CGA | R | GGT | G | CGA | R | GGC | G | CGA | R | GGA | G | CGA | R | GGG | G |
| GGA | G | TGT | C | GGA | G | TGC | C | GGA | G | TGA | . | GGA | G | TGG | W |
| AAA | K | CAT | H | AAA | K | CAC | H | AAA | K | CAA | Q | AAA | K | CAG | Q |
| ACA | T | CCT | P | ACA | T | CCC | P | ACA | T | CCA | P | ACA | T | CCG | P |
| ATA | I | CTT | L | ATA | I | CTC | L | ATA | I | CTA | L | ATA | I | CTG | L |
| AGC | S | CGA | R | AGC | S | CGA | R | AGC | S | CGC | R | AGC | S | CGA | R |
| AGT | S | CGC | R | AGT | S | CGT | R | AGT | S | CGT | R | AGT | S | CGT | R |
| AGG | R | CGG | R | AGG | R | CGG | R |     |   |     |   | AGG | R | CGC | R |

| AGG | | CGT | | AGG | | CGC | | AGG | | CGA | | AGG | | CGG | |
|-----|---|-----|---|-----|---|-----|---|-----|---|-----|---|-----|---|-----|---|
| TGG | W | AGT | S | TGG | W | AGC | S | TGG | W | AGA | S | TGG | W | AGG | S |
| CGG | R | GGT | G | CGG | R | GGC | G | CGG | R | GGA | G | CGG | R | GGG | G |
| GGG | G | TGT | C | GGG | G | TGC | C | GGG | G | TGA | . | GGG | G | TGG | W |
| AAG | K | CAT | H | AAG | K | CAC | H | AAG | K | CAA | Q | AAG | K | CAG | Q |
| ACG | T | CCT | P | ACG | T | CCC | P | ACG | T | CCA | P | ACG | T | CCG | P |
| ATG | M | CTT | L | ATG | M | CTC | L | ATG | M | CTA | L | ATG | M | CTG | L |
| AGA | R | CGA | R | AGA | R | CGA | R | AGA | R | CGC | R | AGA | R | CGA | R |
| AGC | S | CGC | R | AGC | S | CGT | R | AGC | S | CGT | R | AGC | S | CGT | R |
| AGT | S | CGG | R | AGT | S | CGG | R | AGT | S | CGG | R | AGT | S | CGC | R |

Fig. 16 B



## Ser

Fig. 16

## Leu

| TTA | | CTT | | | TTA | | CTC | | | TTA | | CTA | | | TTA | | CTG | |
|-----|---|-----|---|---|-----|---|-----|---|---|-----|---|-----|---|---|-----|---|-----|---|
| ATA | I | ATT | I | | ATA | I | ATC | I | | ATA | I | ATA | I | | ATA | I | ATG | M |
| CTA | L | GTT | V | | CTA | L | GTC | V | | CTA | L | GTA | V | | CTA | L | GTG | V |
| GTA | V | TTT | F | | GTA | V | TTC | F | | GTA | V | TTA | F | | GTA | V | TTG | L |
| TAA | . | CAT | H | | TAA | . | CAC | H | | TAA | . | CAA | H | | TAA | . | CAG | Q |
| TCA | S | CCT | P | | TCA | S | CCC | P | | TCA | S | CCA | P | | TCA | S | CCG | P |
| TGA | . | CGT | R | | TGA | . | CGC | R | | TGA | . | CGA | R | | TGA | . | CGG | R |
| TTG | L | CTA | L | | TTG | L | CTT | L | | TTG | L | CTT | L | | TTG | L | CTT | L |
| TTC | F | CTC | L | | TTC | F | CTA | L | | TTC | F | CTC | L | | TTC | F | CTC | L |
| TTT | F | CTG | L | | TTT | F | CTG | L | | TTT | F | CTG | L | | TTT | F | CTA | L |

| TTG | | CTT | | | TTG | | CTC | | | TTG | | CTA | | | TTG | | CTG | |
|-----|---|-----|---|---|-----|---|-----|---|---|-----|---|-----|---|---|-----|---|-----|---|
| ATG | M | ATT | I | | ATG | M | ATC | I | | ATG | M | ATA | I | | ATG | M | ATG | M |
| CTG | L | GTT | V | | CTG | L | GTC | V | | CTG | L | GTA | V | | CTG | L | GTG | V |
| GTG | V | TTT | F | | GTG | V | TTC | F | | GTG | V | TTA | F | | GTG | V | TTG | L |
| TAG | . | CAT | H | | TAG | . | CAC | H | | TAG | . | CAA | H | | TAG | . | CAG | Q |
| TCG | S | CCT | P | | TCG | S | CCC | P | | TCG | S | CCA | P | | TCG | S | CCG | P |
| TGG | W | CGT | R | | TGG | W | CGC | R | | TGG | W | CGA | R | | TGG | W | CGG | R |
| TTA | I | CTA | L | | TTA | I | CTT | L | | TTA | I | CTT | L | | TTA | I | CTT | L |
| TTC | F | CTC | L | | TTC | F | CTA | L | | TTC | F | CTC | L | | TTC | F | CTC | L |
| TTT | F | CTG | L | | TTT | F | CTG | L | | TTT | F | CTG | L | | TTT | F | CTA | L |

Fig. 17

## Protein Evolution with Expanded Mutation Spectra

Gene to be optimized

Synthesize

Codon usage variant                    Original codon usage

Single
Mutant
Spectrum

accessible          accessible
mutant cloud        mutant cloud
of variant          of original

Ser:  AGY ←→ TCX     can match third base
Arg:  AGR ←→ CGX     for increased homology
Leu:  TTR ←→ TTX

Evolve as a single pool,
shuffle
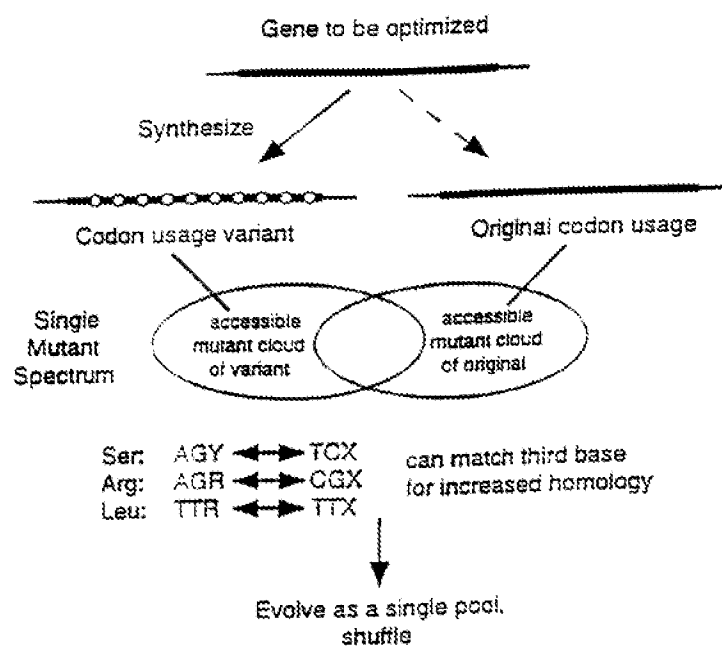
Best synthesis method: all possible codons based on codon synthesis

Fig. 18 A

## Altering the evolution potential of sequences by codon modification

HIV envelope protein (857 amino acids)

Potential mutant resulting from wildtype sequence versus codon modified sequence in arg, leu and serine codons:

Wildtype

Codon-modified

Wildtype

Codon-modified

Fig. 18A

*FIG. 18B FINAL ALIGNMENT*
*NEW ENV VS OLD ENV*    Page 1

Saturday, August 17, 1   δ 2:03 PM
Lipman-Pearson Protein Alignment
kTuple: 2; Gap Penalty: 4; Gap Length Penalty: 12

| Seq1(1>509) new env pro | Seq2(1>509) old env protein | Similarity Index | Gap Number | Gap Length | Consensus Length |
|---|---|---|---|---|---|
| (1>509) | (1>509) | 100.0 | 0 | 0 | 509 |

```
         ▼10        ▼20        ▼30        ▼40        ▼50        ▼60        ▼70
VQYVTVFYGVPAWKNATIPLFCATKNRDTWGTTQCLPDNDDYSELAINVTEAFDAWDNTVTEQAIEDVWN
VQYVTVFYGVPAWKNATIPLFCATKNRDTWGTTQCLPDNDDYSELAINVTEAFDAWDNTVTEQAIEDVWN
VQYVTVFYGVPAWKNATIPLFCATKNRDTWGTTQCLPDNDDYSELAINVTEAFDAWDNTVTEQAIEDVWN
         ▲10        ▲20        ▲30        ▲40        ▲50        ▲60        ▲70
         ▼80        ▼90        ▼100       ▼110       ▼120       ▼130       ▼140
LFETSIKPCVKLTPLCIAMRCNKTETDRWGLTGNAGTTTTAITTTATPSVAENVINESNPCIKNNSCAGL
LFETSIKPCVKLTPLCIAMRCNKTETDRWGLTGNAGTTTTAITTTATPSVAENVINESNPCIKNNSCAGL
LFETSIKPCVKLTPLCIAMRCNKTETDRWGLTGNAGTTTTAITTTATPSVAENVINESNPCIKNNSCAGL
         ▲80        ▲90        ▲100       ▲110       ▲120       ▲130       ▲140
         ▼150       ▼160       ▼170       ▼180       ▼190       ▼200       ▼210
EQEPMIGCKFNMTGLNRDKKKEYNETWYSRDLICEQSANESESKCYMHHCNTSVIQESCDKHYWDAIRFR
EQEPMIGCKFNMTGLNRDKKKEYNETWYSRDLICEQSANESESKCYMHHCNTSVIQESCDKHYWDAIRFR
EQEPMIGCKFNMTGLNRDKKKEYNETWYSRDLICEQSANESESKCYMHHCNTSVIQESCDKHYWDAIRFR
         ▲150       ▲160       ▲170       ▲180       ▲190       ▲200       ▲210
         ▼220       ▼230       ▼240       ▼250       ▼260       ▼270       ▼280
YCAPPGYALLRCNDSNYLGFAPNCSKVVVSSCTRMMETQTSTWFGFNGTRAENRTYIYWHGKSNRTIISL
YCAPPGYALLRCNDSNYLGFAPNCSKVVVSSCTRMMETQTSTWFGFNGTRAENRTYIYWHGKSNRTIISL
YCAPPGYALLRCNDSNYLGFAPNCSKVVVSSCTRMMETQTSTWFGFNGTRAENRTYIYWHGKSNRTIISL
         ▲220       ▲230       ▲240       ▲250       ▲260       ▲270       ▲280
         ▼290       ▼300       ▼310       ▼320       ▼330       ▼340       ▼350
NKYYNLTMRCRRPGNKTVLPVTIMSGLVFHSQPINERPKQAWCWFEGSWKKAIQEVKETLVKHPRYTGTN
NKYYNLTMRCRRPGNKTVLPVTIMSGLVFHSQPINERPKQAWCWFEGSWKKAIQEVKETLVKHPRYTGTN
NKYYNLTMRCRRPGNKTVLPVTIMSGLVFHSQPINERPKQAWCWFEGSWKKAIQEVKETLVKHPRYTGTN
         ▲290       ▲300       ▲310       ▲320       ▲330       ▲340       ▲350
         ▼360       ▼370       ▼380       ▼390       ▼400       ▼410       ▼420
DTRKINLTAPAGGDPEVTFMWTNCRGEFLYCKMNWFLNWVEDRDQKGGRWKQQNRKEQQKKNYVPCHIRQ
DTRKINLTAPAGGDPEVTFMWTNCRGEFLYCKMNWFLNWVEDRDQKGGRWKQQNRKEQQKKNYVPCHIRQ
DTRKINLTAPAGGDPEVTFMWTNCRGEFLYCKMNWFLNWVEDRDQKGGRWKQQNRKEQQKKNYVPCHIRQ
         ▲360       ▲370       ▲380       ▲390       ▲400       ▲410       ▲420
         ▼430       ▼440       ▼450       ▼460       ▼470       ▼480       ▼490
IINTWHKVGKNVYLPPREGDLTCNSTVTSLIAEIDWINSNETNITMSAEVAELYRLELGDYKLIEITPIG
IINTWHKVGKNVYLPPREGDLTCNSTVTSLIAEIDWINSNETNITMSAEVAELYRLELGDYKLIEITPIG
IINTWHKVGKNVYLPPREGDLTCNSTVTSLIAEIDWINSNETNITMSAEVAELYRLELGDYKLIEITPIG
         ▲430       ▲440       ▲450       ▲460       ▲470       ▲480       ▲490
         ▼500
LAPTSVRRYTTTGASRNKR
LAPTSVRRYTTTGASRNKR
LAPTSVRRYTTTGASRNKR
         ▲500
```

DNA HOMOLOGY



Saturday, August 17, 1996 4:39 PM                           Page 1
Wilbur-Lipman DNA Alignment
Ktuple: 3; Gap Penalty: 20; Window: 20

| Seq1(1>1527) new env | Seq2(1>1527) old env | Similarity Index | Gap Number | Gap Length | Consensus Length |
|---|---|---|---|---|---|
| (10>1454) | (10>1454) | (61.7) | 0 | 0 | 1445 |

Fig. 18C

Saturday, August 17, 1996, 4:39 PM
Wilbur-Lipman DNA Alignment
Ktuple: 3; Gap Penalty: 20; Window: 20

| Seq1(1>1527)<br>new env DNA | Seq2(1>1527)<br>old env.pro | Similarity<br>Index | Gap<br>Number | Gap<br>Length | Consensus<br>Length |
|---|---|---|---|---|---|
| (10>1454) | (10>1454) | 61.7 | 0 | 0 | 1445 |

```
        610       620       630       640       650       660
TGGGACGCVATACGTTTCCGCTATTGCGCTCCHCCBGGVTACGCYCTHTTRCGCTGCAAC
TGGGA GC AT  G TT  G TA TG GC CC CC GG TA GC  T  T  G TG AA
TGGGATGCTATTAGATTTAGATACTGTGCACCGCCAGGTTATGCTTTGCTTAGGTGTAAT
        610       620       630       640       650       660

        670       680       690       700       710       720
GACAGTAACTACCTBGGDTTCGCVCCVAATTGCAGYAAAGTTGTCGTAAGCAGYTGTACB
GA    AA TA  T GG TT GC CC AA TG      AA GT GT GT        TG AC
GATTCAAATTATTTAGGCTTTGCTCCTAACTGTTCTAAGGTAGTGGTTTCTTCATGCACA
        670       680       690       700       710       720

        730       740       750       760       770       780
CGBATGATGGAAACCCAGACGAGCACATGGTTCGGDTTTAACGGGVACVCGHGCTGAGAAC
 G ATGATGGA AC CA AC    AC TGGTT GG TT AA GG AC   G GC GA AA
AGAATGATGGAGACGCAAACCTCTACTTGGTTTGGCTTCAATGGTACTAGGGCAGAAAAT
        730       740       750       760       770       780

        790       800       810       820       830       840
CGTACTTATATMTACTGGCACGGDAAGTCVAACCGTACGATTATATCGCTCAACAAATAC
 G AC TA AT TA TGGCA GG AA    AA G AC AT AT     T AA AA TA
AGAACATACATTTATTGGCATGGCAAAAGTAATAGAACCATAATTAGCTTGAATAAGTAT
        790       800       810       820       830       840

        850       860       870       880       890       900
TACAACTTYACTATGCGTTGCCCGTCGCCCTGGBAACAAAACBGTVCTBCCBGTAACGATA
TA AA  T AC ATG G TG  G  G CC GG AA AA AC GT   T CC GT AC AT
TATAATCTAACAATGAGATGTAGAAGACCAGGAAATAAGACAGTTTTACCAGTCACCATT
        850       860       870       880       890       900

        910       920       930       940       950       960
ATGACTGGCCTHGTGTTTCACAGYCAACCDATTAACGAACGTCCTAAGCAAGCDTGGTGT
ATG C GG  T GT TT CA    CA CC AT AA GA  G CC AA CA GC TGGTG
ATGTCAGGGTTGGTCTTCCATTCGCAGCCCATAAATGAGAGACCAAAACAGGCCTGGTGC
        910       920       930       940       950       960

        970       980       990       1000      1010      1020
TGGTTCGAGGGTTCGTGGAAGAAAGCGATACAAGAGGTHAAAGAGACDCTHGTGAAGCAC
TGGTT GA GG    TGGAA AA GC AT CA GA GT AA GA AC  T GT AA CA
TGGTTTGAAGGAAGCTGGAAAAAGGCCATCCAGGAAGTGAAGGAAACCTTGGTCAAACAT
        970       980       990       1000      1010      1020

        1030      1040      1050      1060      1070      1080
CCGCGCTACACHGGBACVAACGACACVCGHAAGATAAACTTGACTGCACCTGCBGGBGGB
CC  G TA AC GG AC AA GA AC  G AA AT AA  T AC GC CC GC GC GG GG
CCCAGGTATACGGGAACTAATGATACTAGGAAAATTAATCTAACAGCTCCAGCAGGAGGA
        1030      1040      1050      1060      1070      1080

        1090      1100      1110      1120      1130      1140
GACCCBGAGGTGACATTCATGTGGACTAACTGCAGRGCBGAGTTTCTBTACTGTAAGATG
GA CC GA GT AC TT ATGTGGAC AA TG  G GG GA TT  T TA TG AA ATG
GATCCAGAAGTCACTTTTATGTGGACAAATTGTCGAGGAGAATTCTTATATTGCAAAATG
        1090      1100      1110      1120      1130      1140

        1150      1160      1170      1180      1190      1200
AACTGGTTCTTRAACTGGGTBGAAGATCGBGATCAGAAAGGAGGGGCGTTGGAAGCAGCAG
AA TGGTT   AA TGGGT GA GA  G GA CA AA GG GG  G TGGAA CA CA
AATTGGTTTCTTAATTGGGTAGAGGACAGAGACCAAAAGGGTGGCAGATGCAAACAACAA
        1150      1160      1170      1180      1190      1200
```

Fig 18C

Saturday, August 17, 1   , 4:39 PM
Wilbur-Lipman DNA Alignment
Ktuple: 3; Gap Penalty: 20; Window: 20

| Seq1(1>1527)<br>new env DNA | Seq2(1>1527)<br>old env.pro | Similarity<br>Index | Gap<br>Number | Gap<br>Length | Consensus<br>Length |
|---|---|---|---|---|---|
| (10>1454) | (10>1454) | 61.7 | 0 | 0 | 1445 |

```
   ↗1210    ↗1220     ↗1230    ↗1240     ↗1250     ↗1260
AACCGHAAGGAACAGCAAAAAAAGAACTACGTHCCBTGCCACATMCGBCAGATTATAAAT
AA  G AA GA CA CA AA AA AA TA GT CC TG CA AT  G CA AT AT AA
AATAGGAAAGAGCAACAGAAGAAAAATTATGTGCCATGTCATATTAGACAAATAATCAAC
 ↖1210    ↖1220     ↖1230    ↖1240     ↖1250     ↖1260
   ↗1270    ↗1280     ↗1290    ↗1300     ↗1310     ↗1320
ACCTGGCATAAGGTBGGDAAGAACGT3TACCTCCCACCCCGCGAGGGTGATTTRACBTGT
AC TGGCA AA GT GG AA AA GT TA  T CC CC  G GA GG GA  T AC TG
ACGTGGCACAAAGTAGGCAAAAAFGTATATTTGCCTCCTAGGGAAGGAGACCTGACATGC
 ↖1270    ↖1280     ↖1290    ↖1300     ↖1310     ↖1320
   ↗1330    ↗1340     ↗1350    ↗1360     ↗1370     ↗1380
AACAGYACVGTTACATCATTGATTGCTGAAATYGACTGGATWAACTCDAACGAAACGAAC
AA    AC GT AC     T AT GC GA AT GA TGGAT AA      AA GA AC AA
AATTCCACTGTAACTAGTCTCATAGCAGAGATAGATTGGATCAATAGCAATGAGACCAAT
 ↖1330    ↖1340     ↖1350    ↖1360     ↖1370     ↖1380
   ↗1390    ↗1400     ↗1410    ↗1420     ↗1430     ↗1440
ATAACGATGTCVGCBGAAGTHGCBGAGTTYTACAGGCTCGAACTCGGTGACTATAAGCTB
AT AC ATG   GC GA GT GC GA  T TA  G  T GA  T GG GA TA AA  T
ATCACCATGAGTGCAGAGGTGGCAGAACTGTATCGATTGGAGTTGGGAGATTACAAATTA
 ↖1390    ↖1400     ↖1410    ↖1420     ↖1430     ↖1440
   ↗1450
ATYGA
AT GA
ATAGA
 ↖1450
```

Saturday, August 17, 1996 ∠ ⅃ PM          F₃        18ℓ                    Page 1
1.6 kb Env mutagenized
Created: Friday, August 16, 1996 9:39 AM

MUTAGENIZED GENE
IN CONTEXT

```
          10        20        30        40
   |    |    |    |    |    |    |    |    |    |
aactccgaagaagactaaggctaatccatttcctgcatca 40
aacaagtaagtatgggatgtcttgggaatcagctgcttat 80
cgcgatcttgctagtaagtgttttagagatttgttgtGTv 120
CAgTAcGTbACbGTbTTtTAcGGaGTtCCTGCtTGGAAaA 160
AcGChACbATmCCattrTTtTGcGCtACaAAaAAccgnGA 200
         210       220       230       240
   |    |    |    |    |    |    |    |    |    |
tACvTGGGbACbACbCAgTGtctcCCtGAcAAcCAcGAc 240
TAtagyGAgcthGCbATaAAcGTaACtGAaGCaTTcGAcG 280
CvTGGGAcAAcACbGTdACtGAgCAgGCtATtGAaGAcGT 320
hTGGAAtttrTTcGAgACtagtATaAAcCCaTGcGTtAAg 360
ttrACdCCbttgTGcATtGCtATGAGgTGcAAcAAgACvG 400
         410       420       430       440
   |    |    |    |    |    |    |    |    |    |
AaACbGAccgnTGGGGvctcACtGGtAAtGCtGGtACbAC 440
bACbACbGCbATyACbACtACcGCtACbCCttcvGTtGCb 480
GAgAAcGTyATyAAcGAgtcaAAcCCaTGtATtAAgAAcA 520
ActcvTGcGCbGGdctyGAgCAcGAaCCdATGATtGGvTG 560
cAAgTTcAAtATGACbGGccttAAccgcGAtAAgAAgAAg 600
         610       620       630       640
   |    |    |    |    |    |    |    |    |    |
GAgTAcAAcGAgACbTGGTAcagtcgtGAccttATaTGcG 640
AaCAaagyGChAAcGAgtcaGAatccAAgTGcTAtATGCA 680
cCAcTGcAAtACdtcvGTaATaCAgGAgagtTGcGAtAAa 720
CAcTAcTGGGAcGCvATacgtTTccgcTAtTGcGCtCChC 760
CbGGvTAcGCvcthttrcgcTGcAAcGAcagtAAcTAcct 800
         810       820       830       840
   |    |    |    |    |    |    |    |    |    |
bGGdTTcGCvCCvAAtTGcagyAAaGTtGTcGTaagcacy 840
TGtACbcgbATGATGGAaAccCAgACgagcACaTGGTTcG 880
GdTTtAAcGGvACvcghGCtGAgAAccgtACtTAtATmTA 920
cTGGCAcGGdAAgtcvAAccgtACgATtATatcgctcAAc 960
AAaTAcTAcAActtgACtATGcgtTGccgtcgcCCtGGbA 1000
        1010      1020      1030      1040
   |    |    |    |    |    |    |    |    |    |
AcAAaACbGTvctbCCbGTaACgATaATGagtGGccthGT 1040
dTTtCAcagyCAaCCdATtAAcGAacgtCCtAAgCAaGCd 1080
TGGTGtTGGTTcGAgGGttcgTGGAagAAaGCgATaCAaG 1120
AgGThAAaGAgACdcthGTgAAgCAcCCgcgcTAcAChGG 1160
bACvAAcGAcACvcgnAAgATaAActtgACtGCaCCtGCb 1200
```

Figure 19

ENV Top oligos

1T CTAGTAAGTGTTTTTGAGAGTTGTTGT

2T TTACGGAGTTCCTGCTTGGAAAAACGCTACTATTXCCTTRTTTTTGCGCTACAAAAAACCG

3T AGTGTGTCTCCCTGTCAACGACGACTATTAGYGTAGCTTGCTATAAACGTTAACTGAAGCATTCG

4T ACTGAGCTGGCTATTGAAGACGTATTGGAATTTRTTCGAGCTAGTATAAAAACCATGCGTT

5T TGCTATGAGGTGCAACAAGACVGAAACBGACCGTTGGGGVCTCACTGGTAATGCTGGTAC

6T CTACCGCTACTCCTTCAGTTGCBGAGACGTTATYAACGAGTCAAACCCTTGTTTTAAGA

7T GAGCAAGAACCGATGATTGGVTGCAAGTTCAATATGACBGGCCTTAACCGCGTATAAGAAA

8T GTACAGTCGTGACCTTATATGCGAACAAAGTGCTAACGAGTCAGAATCCAGTGCTATAT

9T TAATACAGGAGAGTTCGGATAAACACTACTGCGAGCGCTATACGTTTCCCCTATTGCGCTC

10T CGCTGCAACGACAGTAACTACCTGGGTTCGCVCCVAATTGCAGTAAAGTTGTCGTAAGC

11T AACCCAGACGAGCCACATGGTTCGGDTTTAACGGVACVCGTAGCTGAGAACCGTACTTATAT

12T ACCGTACGGATTATATCGCTCAACAAAATACTACAACTTGACTATGCGTTGCCGTCGCCCTG

13T GTAACGATAATGAGTCGCCTTGTTDTTTCACAGYCAACCTATTAACGAACGTCCTAAGCAA

14T TTCGTGGAGGAAGCCGATACAAGAGGTHAAAGAGACDCTTGTTGAAGCACCCCTCGCTCCAC

15T AGATAAACTTGACTGCACCTGCBGGBGGBGACCCBGAGGTGACATTCATGTGGACTAACT

16T TGTAAGATGAACTGGTTCTTTAACTGGGTBGAAGATTGBGATCAGAAAGGAGGGCGTTGG

17T ACAGCAAAAAAAGAACTACGTTCCCBTGCCTACATTXCBCAGATTATTAAATACCTGGCATAA

18T TCCCTACCCCCGCGTAGGGTGATTTTRACBTGTTAACAGTACVGTTTACATCATTGATTCCTGAAA

19T GAAACGAACATAACGTTGTCVGCBCGAGTTGCCBGAGTTRTTACACGGCTCGAACTCGGTGAC

20T ACCTTATAGGGCTCGCGCCCGACBTCVGTBCCBCCGTTATACTACTACAGGTGCGGAGTCGTAA

Fig 19 (cont.)

ENV BOTTOM OLIGOS

1B   CCCCVCGTTTGTTACGACTCGCTCCTGTAGT

2B   CTGGCGCGAGCCCTATAGGTGTKATTTCRATVAGCTTATAGTCACCGAGTTCGAGCCTGT

3B   GACATCGTTATGTTCGTTTCGTTKGAGTTWATCCAGTTCRATTTCAGCAATCANTGATGTA

4B   ATCACCCTCCCGCGGTGGCAGTAVACGTTCTTRCCVACCTTATGCCAGGTATTTATAAT

5B   CGTAGTTCTTTTTTGCTGTTCCTTTCGGTTCRCCTGCTTCCAACGCCCTCCTTTCTGAT

6B   AAGAACCAGTTCATCTTTCAGTAVAGRAACTCVCCXCTGCAGTTAGTCCACRTGAATGTC

7B   AGGTGCRGTCAAGTTTRTCTTCCGRGTGTCGGTTEGTVCCCGTGTGTAGCGCGGGTGCTTCAC

8B   GTATCGCTTTCTTCCRCGRACCCTCGAACCAACRCCRRGCTTGCTTAGGRCGTTCGTTRA

9B   AGGCCACTCRTTATCGTTRCVGGVAGRACVGTTTTGTTVCCRGGGCGACGGCTAACTCATA

10B  GRGCGRTATRATCGTRCGGTTBGACTTBCCGTGCCAGTAKATATRAGTACGGTTCTCAGC

11B  ACCATGTGCTCGTCTGGGTTTCATCATVCGVGTACARCTGCTTACGACAACTTTACTGC

12B  TRGTTACTGTCGGTTCCRGCGTAADAGRGCGGTRBCCVCGCGGAGCGCRATAGCGGRAACGT

13B  ATCGCRACTCTCCTGTATTACBCARGTATTGCAGTGGTGCATATRGCRCTTGGRTTCTCGR

14B  ATRTRAGGTCRCGACTGTRCCRVGTCTCGTTGTRCTCCTTTTTTCTTRTCGCGGTTRAGGC

15B  CCRATCRTCGGTTCTTCGCTCRRGRCCVGCGCRRGCRGTTGTTCTTRRTRCRTGGGTTTGRC

16B  AACTGRRGGRGTRGCGGTRGTVGTRRTVGCVGTVGTVGTVGTRCCRGCRTTRCCRGTGRG

17B  TCTTGTTGCRCCTCRTRGCRRTGCRCRRVGGRBGTYRRCTTRRCGCRTGGTTTTRTRCTRG

18B  TCTTCRRTRGCCTGCTCRGTRRCVGTGTTGTCCCRBGCGTCGRRTGCTTCRGTTRCGTTT

19B  GTCGTTGTCRGGGRGRCRCTGVGTVGTVCCCRBGTRTCTCGGTTTTTTGTRGCGCRRRR

20B  TCCRRGCRGGRRCTCCGTRRRRVRCVGTVRCGTRCTGRCRCRRCRRRTCTCTRRRRCRC